Universidad Nacional Mayor de San Marcos

Universidad del Perú. Decana de América

Facultad de Ingeniería de Sistemas e Informática

Escuela Académico Profesional de Ingeniería de Sistemas

Aplicación de la minería de datos para la optimización del proceso de presupuestación usando algoritmos de serie temporal

TESINA

Para optar el Título Profesional de Ingeniero de Sistemas

AUTORES

Diana Su WING LENT

Carlos Christian VALERIO ORDÓÑEZ

ASESOR

Percy Edwin DE LA CRUZ VÉLEZ DE VILLA

Lima, Perú

2009

Referencia bibliográfica

Wing, D. & Valerio, C. (2009). *Aplicación de la minería de datos para la optimización del proceso de presupuestación usando algoritmos de serie temporal*. Tesina para optar grado el título profesional de Ingeniero de Sistemas. Escuela Académico Profesional de Ingeniería de Sistemas, Facultad de Ingeniería de Sistemas e Informática, Universidad Nacional Mayor de San Marcos, Lima, Perú.

Dedicamos el presente trabajo a nuestros padres que siempre confiaron en nosotros y a nuestra hija Adriana que ha sido nuestra mayor motivación para salir adelante.

Ficha Catalográfica

Autores:

Diana Su Wing Lent
Carlos Christian Valerio Ordoñez

Título de tesina:

Aplicación de la Minería de Datos para la Optimización del Proceso de Presupuestación usando Algoritmos de Serie Temporal

Nombre de línea de investigación:

Ingeniería del Conocimiento

Lima 2010

UNMSM, Pregrado, Ingeniería de Sistemas e Informática

Tesina, Universidad Nacional Mayor de San Marcos, Facultad de Ingeniería de Sistemas e Informática

RESUMEN

El presente trabajo tiene como objetivo presentar una solución de Inteligencia de Negocios para optimizar el proceso de Presupuestación en Operaciones Mineras usando Series Temporales como técnicas de Minería de Datos.

La minería de datos nos permite encontrar nuevos conocimientos a partir de ciertos patrones y modelos, los cuales se basan en información histórica que residen en un almacén de datos. De esta manera, se decidió automatizar el análisis estadístico y las proyecciones de costos basadas en datos históricos. Además, se integraron las diversas fuentes de datos en un repositorio de datos, de donde se obtienen los ratios de los principales consumibles, con la finalidad de reducir el tiempo total del proceso de presupuestación y mejorar la calidad y confiabilidad de los costos proyectados.

El uso de series temporales ha permitido obtener resultados más cercanos a la realidad de la organización, en este caso al área de Operaciones Mina, por lo que se recomienda su aplicación y adaptación a las demás áreas operativas, entre ellas Procesos y Mantenimiento.

Palabras Claves: Costos, Minería de Datos, Presupuestos, Series Temporales.

ABSTRACT

This study aims to provide a Business Intelligence solution for optimize the process of Budgeting in Mining Operations using the technique of Time Series in Data Mining.

The data mining allows us to find new knowledges from certain patterns and models, which are based on historical information that reside in a database. In this way, we decided to automate the statistical analysis and the costs projections based on historical information. In addition, the diverse data sources were joined in a repository of data, where we get the ratios of the principal consumables, in order to reduce the total time of the process of Budgeting and improve the quality and reliability of the projected costs.

The use of time series has allowed to obtain results nearer to the reality of the organization, in this case the Mine Operations area, for that reason, we recommend its application and adjustment to other operative areas, as Process and Maintenance areas.

Key Words: Costs, Data Mining, Budget, Time Series.

TABLA DE CONTENIDO

Lista de Figuras

Lista de Tablas

CAPÍTULO 1: Introducción

Hoy en día no hay nada nuevo en el mundo, sino que el esfuerzo de cada ser humano es la prolongación de lo que cada generación anterior ha logrado, y el mérito de cada hombre, cada profesional está en su aporte aunque sea casi insignificante, pero que será un eslabón más de todo el conocimiento que aporta el hombre y ayuda al progreso de toda la humanidad.

En la investigación plasmaremos nuestros conocimientos y experiencias dentro de nuestra línea de carrera: Ingeniería, con la finalidad de solucionar un problema de la sociedad que nos rodea, aplicando los métodos y tecnologías necesarias y de esta manera aportar un nuevo conocimiento.

El problema a abordar está relacionado con la optimización del proceso de presupuestación en el área de Operaciones Mina del emplazamiento minero Pierina, ubicado en Huaraz y perteneciente a la Compañía Minera Barrick Misquichilca; se busca minimizar el tiempo empleado en el proceso mencionado, integrar los datos históricos almacenados en distintas bases de datos y mejorar la confiabilidad de los datos usados como variables de entrada para obtener el costo total de los principales consumibles del Presupuesto del área.

Actualmente sabemos que, gracias a las TIC (Tecnologías de Información y Comunicación) disponibles en el mercado, es posible almacenar, organizar, procesar y extraer información valiosa a partir de grandes cantidades de datos. Es necesario entonces contar con un sistema que juegue el papel de soporte para la toma de decisiones, de respuesta ágil y rápida, con información precisa para poder aprovechar las oportunidades: *"estar en el lugar indicado, en el momento oportuno, con la información correcta"*, optimizando de esta manera los procesos y mejorar la productividad.

Por lo tanto, el presente trabajo tiene como objetivo principal presentar una solución tecnológica que ayude a optimizar el proceso de presupuestación dentro de una organización, que para este caso será el área de Operaciones Mina de una empresa minera. Para lograr ello, aplicaremos series temporales, una solución de la Minería de Datos que se encuentra dentro de las técnicas aplicadas en Inteligencia de Negocios, la cual brinda un análisis predictivo de datos basándose en información histórica.

La aplicación de esta solución permitirá ahorrar esfuerzos y tiempo a los ejecutivos encargados de la elaboración del presupuesto, además de contar con una mejor calidad de información.

1.1 Planteamiento del Problema

En esta sección se define el problema, haciendo una descripción de su entorno actual, sus antecedentes y terminamos formulando el problema en forma de interrogantes.

1.1.1 Descripción de la realidad problemática

En la actualidad, la minería juega un rol importante en la economía peruana y más aún en las economías regionales donde se encuentran los asentamientos mineros.

En el ámbito nacional, la importancia de la minería cobra relevancia sobre todo a nivel de los recursos adicionales que le genera a la economía a través de flujos de capital provenientes de inversiones y exportaciones.

Como podemos observar en la Figura 1.1, en los últimos siete años el sector minero ha experimentado un importante dinamismo y desarrollo aumentando de manera espectacular tanto los recursos por divisas generados como los impuestos que paga. Es decir, el proceso de expansión ha significado también un incremento de la importancia de la minería en la economía nacional.

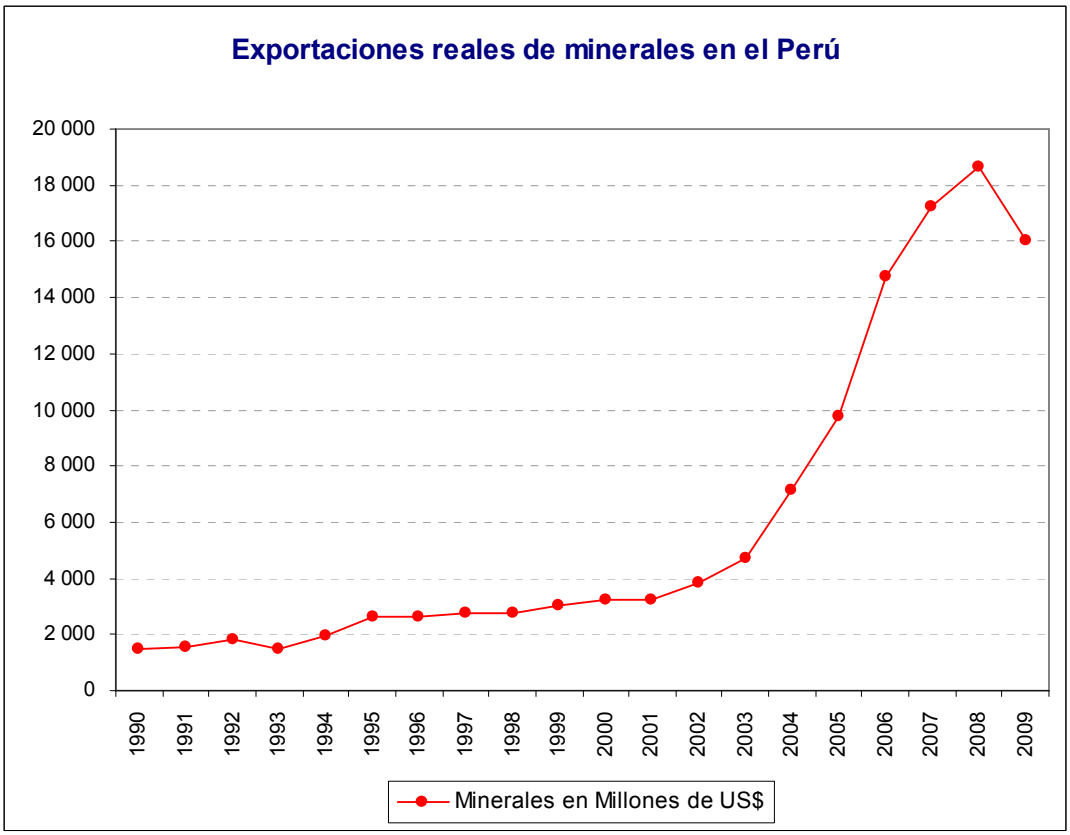


Figura 1.1 Evolución de las exportaciones mineras en el Perú. Fuente: [MEM01]

Esta evolución de la minería en el Perú, trae consigo una mayor cantidad de procesos operativos en las empresas, los cuales conllevan un aumento de costos que deben ser bien presupuestados con la finalidad de obtener una mayor productividad.

En la elaboración de un presupuesto influyen muchos indicadores, los cuales, si no son bien calculados, pueden llevar a un mal presupuesto y esto a su vez generar grandes pérdidas para la empresa.

Para el caso de las empresas mineras, como es bien sabido que manejan grandes cantidades de dinero, se deben buscar todos los medios posibles que ayuden a optimizar el proceso de presupuestación.

Por esta razón, es que nos planteamos la elaboración de una aplicación de minería de datos que ayude a mejorar el proceso realizado por los analistas de costos para obtener un presupuesto rentable y confiable.

1.1.2 Antecedentes del Problema

Por muchos años, la minería en Perú ha sido y viene siendo uno de los pilares del desarrollo económico del país. Las empresas mineras han sido y seguirán siendo protagonistas en este proceso.

El aumento de precios de los metales produce mayores utilidades, pero también trae consigo alzas en los precios de los insumos. Por ello, es importante cuidar los costos correspondientes a estos.

La gestión de costos es fundamental en la operación, generando continuas iniciativas para mejorar la eficiencia del proceso de presupuestación.

Las relaciones entre consumos y factores causales de los costos son la base del modelo, permiten proyectar gastos a otros períodos, analizar los costos actuales y comparar con datos históricos.

El proceso de presupuestación de costos variables, el cual se aplica en el área de Operaciones Mina, está basado en indicadores de consumo por ratios que se obtienen de análisis estadístico de data histórica.

Por lo tanto el proceso de elaboración de un presupuesto puede demandar varias horas, días y hasta semanas desde el inicio hasta su aprobación final por la Gerencia General.

Este proceso puede, muchas veces, causar estrés y conflictos entre los analistas de costos debido a que tienen que procesar una gran cantidad de datos y basarse en información histórica para realizar la proyección de sus gastos futuros.

El problema surge debido a que muchas veces esta data histórica se encuentra en diferentes fuentes de almacenamiento de datos.

En ciertas ocasiones, los analistas de costos, tienen que valerse de su propia “intuición” y calcular ciertos gastos en base a su propia experiencia en el área.

1.1.3 Formulación del Problema

El proceso de presupuestación en Operaciones Mina está basado en indicadores de consumo por ratios para los principales insumos, por ejemplo tenemos que: para el cálculo de diesel (petróleo) usado en los equipos mineros es necesario realizar un análisis estadístico para obtener un indicador de galones por hora, esto se obtiene de diversas fuentes y tarda varios días el lograr consolidarse, ya que no cuenta con un repositorio de datos integrado y mucho menos actualizado.

Actualmente, los datos provienen de diversas fuentes como hojas de cálculo de Excel, base de datos Access y Oracle, estos datos no se encuentran clasificados para un fácil y rápido análisis estadístico de los mismos, lo cual demanda tiempo al personal encargado de presupuestar en base a estos indicadores. Por lo tanto podríamos hacernos la siguiente interrogante:

¿Cómo se podría reducir el tiempo de obtención de los ratios de los consumibles para optimizar el proceso de presupuestación?

Por otro lado, se manejan varios indicadores de costos unitarios según la producción del área, se usa el costo por tonelada minada para el caso de Operaciones Mina, el costo por onza producida para el caso de Procesos, adicionalmente se tienen costos fijos y semivariables para las áreas administrativas. A partir de esta problemática nos hacemos la siguiente pregunta:

¿En qué medida se puede automatizar el análisis estadístico de datos históricos para aumentar la confiabilidad de los índices de consumibles?

El problema encontrado en este proceso de presupuestación es que los analistas de costos deben recurrir a los datos provenientes de diversas fuentes de datos y luego realizar un análisis estadístico para obtener los ratios de consumo de los principales insumos en Operaciones Mina; lo que demanda, en el mejor de los casos, varias horas de trabajo cuando los datos están accesibles pero deben analizarse; en el peor de los casos, este proceso puede tardar varios días cuando los datos requeridos no están almacenados en las fuentes de datos establecidas.

En este contexto, se busca desarrollar una aplicación que posibilite optimizar esta parte inicial en el proceso de presupuestación, integrando en un repositorio los

datos provenientes de diversas fuentes y automatizando el cálculo de los índices de cantidades de los consumibles, con la finalidad de obtener un presupuesto confiable, rentable y haciendo uso de una menor cantidad de horas-hombre.

1.2 Objetivos

Dentro del presente trabajo de investigación nos hemos propuesto los siguientes objetivos:

1.2.1 Objetivo principal

- Desarrollar una solución de Inteligencia de Negocios para optimizar el proceso de presupuestación en operaciones mineras a través del uso de la Minería de Datos.

1.2.2 Objetivos secundarios

- Automatizar al máximo el análisis estadístico y proyecciones basadas en datos históricos para aumentar la confiabilidad de los índices de consumibles.

- Integrar las diversas fuentes de datos en un repositorio de donde se obtendrán los ratios de los consumibles para mejorar el proceso de presupuestación.

1.3 Limitaciones de la Investigación

En esta sección limitaremos nuestro trabajo de investigación desde la perspectiva espacial, social y técnica.

1.3.1 Delimitación Espacial

El presente trabajo se delimita al estudio de la problemática del proceso de Presupuestación en el Área de Operaciones Mina de la empresa Minera Barrick Misquichilca S.A..

1.3.2 Delimitación Social

Este estudio se delimita al beneficio que puedan conseguir los ejecutivos encargados de elaborar el presupuesto cada año, puesto que se minimizará el tiempo de presupuestación y esto traerá consigo una mejor calidad de vida para ellos. Además del beneficio que pueden obtener las personas encargadas de tomar decisiones en base a un presupuesto adecuado, trayendo consigo mayores beneficios para la organización.

1.3.3 Delimitación Técnica

Esta investigación delimita su estudio técnico al análisis, diseño e implementación de una solución de minería de datos para optimizar el proceso de Presupuestación en el área de Operaciones Mina de la empresa Minera Barrick Misquichilca S.A. Esta solución capturará los datos almacenados en diferentes fuentes tales como archivos planos, Excel y tablas de base de datos, para luego almacenarlos en un modelo de base de datos multidimensional, el paso siguiente será procesar esta información, encontrar ciertos patrones y en base a estos, proyectar indicadores clave para la elaboración del presupuesto.

1.4 Justificación

La influencia de la minería en la economía nacional es evidente. Para ello, basta reconocer su importancia en la generación de recursos adicionales (exportaciones, impuestos, inversiones) los que tienden a financiar parte de la estabilidad económica del país y su proceso de desarrollo.

Para graficar la importancia del sector minero en la economía peruana se muestra la Figura 1.2, que representa la estructura porcentual de las exportaciones en el Perú durante los últimos 20 años. Como observamos, en los últimos 5 años las exportaciones provenientes de la minería representan el 60% del total de exportaciones en el Perú, con respecto al sector pesquero, petróleo, agrícola y otros productos no tradicionales.

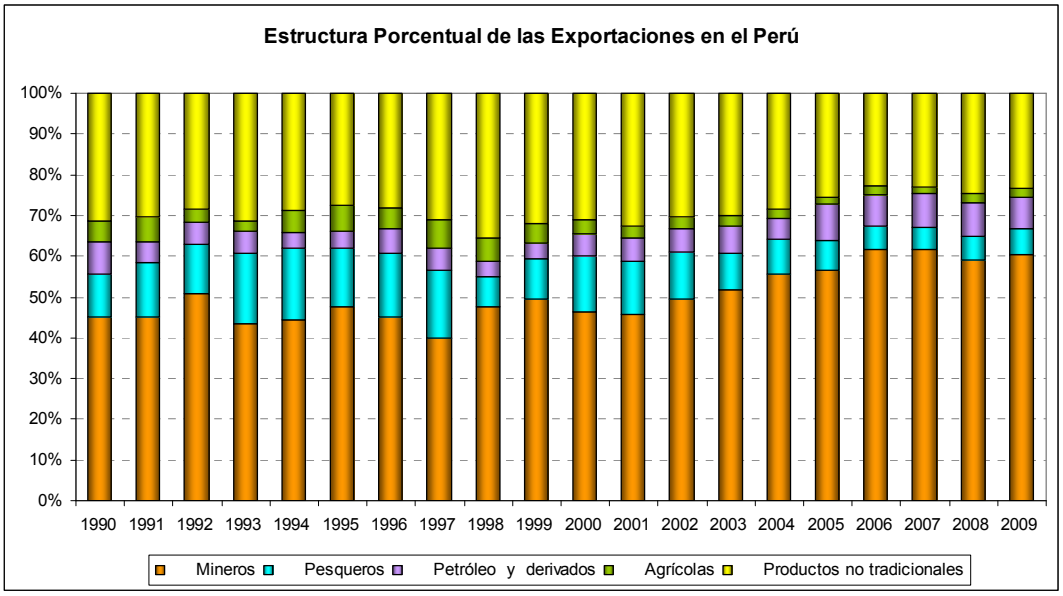


Figura 1.2 Estructura Porcentual de las Exportaciones en el Perú. Fuente [MEM01]

Tomando en cuenta la importancia del rol protagónico de las empresas mineras en el desarrollo de nuestro país, podemos encontrar que una de las principales áreas de estas empresas es la de Operaciones Mina, por ser la encargada de la parte operativa. Esta área a su vez realiza anualmente un presupuesto "Budget" (proceso mediante el cual los directivos de una empresa aseguran la utilización eficaz y eficiente de los recursos en el futuro) donde se proyectan los gastos a realizarse en el período futuro, adicionalmente se realizan actualizaciones a este presupuesto, cada dos o tres meses, a los que denominan "Forecast" (presupuesto conformado por datos reales y datos proyectados).

La elaboración de un presupuesto rentable y confiable es vital para la correcta operatividad de la empresa, ya que se manejan grandes cantidades de dinero.

Para el desarrollo del presupuesto, los analistas de costos tienen que basarse en grandes cantidades de información histórica, lo cual servirá de guía para proyectar los gastos a futuro.

Es por ello, que es muy importante contar con una solución que permita manejar de manera óptima los datos que servirán para el desarrollo del presupuesto dentro de estas empresas, tomando en cuenta que en la mayor parte de las organizaciones, los procesos se alojan en múltiples sistemas, distribuidos por toda la empresa. Los ejecutivos encargados de este presupuesto en las organizaciones tienen que hacer frente inevitablemente al reto de procesar gran cantidad de datos. Por consiguiente,

los procesos de inteligencia de negocios o "business intelligence" (BI), suelen ser una estrategia bastante práctica para las organizaciones que se preocupan de la racionalización de sus procesos para mejorar sus resultados comerciales. Y dentro de esta tecnología BI ubicamos la herramienta de Minería de Datos como la mejor opción para optimizar este proceso de Presupuestación, el cual se basa en data histórica para lograr la prospección de datos, lo cual nos servirá para poder proyectar costos al momento de elaborar el presupuesto.

Por lo tanto, el presente trabajo se justifica bajo las siguientes premisas:

a. La automatización del análisis estadístico de los datos históricos aumentará la confiabilidad de los índices de los consumibles, mejorando la calidad de los datos.

b. La reducción del tiempo de obtención de ratios de los consumibles permitirá optimizar el proceso de presupuestación, minimizando los tiempos.

c. La integración de las diversas fuentes de información en un repositorio de datos ayudará a automatizar el análisis estadístico de los datos.

1.5 Propuesta

Dentro de la iniciativa tecnológica propuesta, encontramos la Inteligencia de Negocios como una opción de Tecnologías de Información, la cual cuenta con una suite de Soluciones, Tecnologías y Metodologías diversas para resolver problemáticas muy complejas de Reporting, Análisis, Minería de Datos (Data Mining), Gestión del Conocimiento y Gestión del Rendimiento, además de herramientas de usuario final como 'ad-hoc query', 'analysis and reporting' -incluyendo tableros de mando- y, finalmente, la producción de informes a partir de cualquier clase de datos de empresa.

Dado que se parte de un conjunto de datos históricos, se plantea utilizar una metodología orientada a la minería de datos, ya que se pretende conseguir mediante técnicas y herramientas extraer un conocimiento implícito, que actualmente no conocemos y se encuentra almacenado en el conjunto de datos. Utilizar esta metodología tiene como objetivo predecir de forma automatizada tendencias y comportamientos o construir un modelo desconocido.

Nuestra propuesta incluye la implementación de una Data Warehouse como un paso previo a la implementación final de la Minería de Datos (Data Mining) para la proyección de indicadores basados en data histórica lo cual va a permitir a los usuarios poder obtener, en base a información histórica, y mediante analítica avanzada, los indicadores necesarios para desarrollar su presupuesto óptimamente, ahorrándose tiempo en el análisis al tener la información integrada y calculada.

La técnica usada para la prospección de los datos será el algoritmo de series temporales, el cual nos permitirá obtener un modelo a partir de los datos históricos residentes en el Data Warehouse. Tomando en cuenta que el conocimiento es uno de los activos más valiosos de una empresa, se va a desarrollar una solución para recopilarlo, analizarlo y proyectarlo en el tiempo.

A continuación pasamos a detallar más específicamente lo que incluirá nuestra solución propuesta:

- El Data Warehouse es un sistema informático centralizado e integrado de almacenamiento de información, el cual está orientado a la generación de informes para el análisis por parte de usuarios finales. La información puede obtenerse de los distintos sistemas existentes en cada departamento de la organización, pudiendo abarcar cualquier temática sin limitación de detalle, volumen o antigüedad. Se trata del almacén o repositorio de información corporativa, derivado directamente de los sistemas operacionales y de orígenes de datos externos, organizado por áreas de interés y no volátil, que contiene la información para apoyar el proceso de toma de decisiones de una organización.

- La solución de Minería de Datos, basada en Inteligencia de Negocios es una aplicación informática que consolida y analiza datos de negocio aún no procesados y los convierte en información concluyente y manejable. Permite a las empresas tener acceso y recopilar información de diferentes fuentes como datos de clientes, de operaciones y de mercado, para obtener ventajas competitivas de su uso. Aporta a las compañías la inteligencia necesaria para identificar tendencias, mejorar relaciones, reducir riesgos financieros y crear nuevas oportunidades de mejora.

1.6 Organización de la Tesina

El primer capítulo explica el objeto de estudio del presente trabajo, brindando información acerca del problema y la solución propuesta.

El resto del documento se organiza de la siguiente manera. En el segundo capítulo se realiza la definición de un proceso de optimización, una explicación del proceso de presupuestación minera y sus conceptos, así como las variables que se incluyen, luego se definirá el concepto de Minería de Datos, y se concluye este capítulo con la definición del concepto de Series Temporales. El tercer capítulo abarca el Estado del Arte donde mostraremos las herramientas existentes en el mercado para la implementación de una solución basada en un modelo de minería de datos y las metodologías de desarrollo de proyectos de minería de datos. En el cuarto capítulo se describe el aporte teórico de la solución, eligiendo a las series temporales y la metodología CRISP-DM para la implementación de la presente aplicación. En el quinto capítulo, el aporte práctico, se describirán los pasos para implementar la solución elegida al problema objeto de estudio. En el sexto capítulo mostraremos el desarrollo de la solución tecnológica, con el diseño, análisis e implementación del sistema. En el séptimo capítulo desarrollaremos el análisis y recolección de datos a fin de validar la solución propuesta. En el octavo capítulo indicaremos nuestras conclusiones y opinión acerca de los futuros trabajos a ser considerados. En el noveno capítulo se muestran las referencias bibliográficas. Finalmente, en el Anexo A mostraremos el caso de estudio, ubicando el área de la empresa benficiada en el organigrama de la misma, así como también se muestra un diagrama de la situación actual del proceso de presupuestación y un diagrama con la solución propuesta en el presente trabajo aplicativo.

CAPÍTULO 2: Marco Teórico

En el presente capítulo se establecen las bases teóricas del proyecto al igual que el conocimiento previo que hay que tener en cuenta para el desarrollo del mismo. Trataremos algunos temas relacionados con el proceso de elaboración de presupuestos dentro una organización. Asimismo, abordaremos algunos conceptos de minería de datos, los cuales ayudarán a entender mejor el tema del presente trabajo de investigación.

2.1 Optimización de Procesos

Este término se compone de dos conceptos claves: Optimizar y Proceso, por lo tanto empezaremos antes por definir ambos términos:

2.2.1 Procesos

Un proceso es un conjunto de acciones y actividades interrelacionadas que se llevan a cabo para alcanzar un conjunto previamente especificado de productos, resultados o servicios. [PMBOK04]

Esta definición puede ser complementada con la Norma ISO 10006, la cual agrega: *"Un Proceso es el conjunto de actividades mutuamente relacionadas o que interactúan, las cuales transforman elementos de entrada en resultados"*.

2.2.2 Optimizar

Es encontrar el mínimo o el máximo de una función respecto a ciertas restricciones. Sin duda, alcanzar el mínimo o máximo es obtener la "mejor" solución

entre otras soluciones factibles. Ahora bien, el mejor proceso debe ajustar el flujo de tareas, entradas y salidas de manera que entregue la mejor calidad al menor costo y en el menor tiempo. Sin embargo, si queremos aumentar la calidad de un producto o servicio (core process) siempre se incurre en inversión de tecnología y personas (costos aumentan) pero a la vez se pueden reducir los tiempos (de producción, soporte, time-to-market, etc.) o bien aumentarlos, esto depende de otros factores tales como: correcta elección de la tecnología, capacitación de las personas, estrategias de gestión (gestión del cambio y gestión del conocimiento).

Alternativamente, si queremos reducir los costos asociados al producto o servicio (core process) muchas veces las empresas disminuyen los tiempos pero a la vez disminuye la calidad. De este modo, si queremos reducir los tiempos asociados al producto o servicio una vez más incurrimos en costos y reducción de la calidad. Finalmente, la flexibilidad de un proceso está asociada a cuán rápido se ajusta a los cambios y dinamismo de la empresa y del entorno, los cuales podemos dividir en factores internos y externos.

Los factores internos son aquellas medidas e iniciativas de la empresa para realizar cambios a un proceso para mejorar su desempeño tomando en cuenta las variables de costo, tiempo, calidad y flexibilidad. Los factores externos son todos aquellos factores que provienen desde el entorno de la empresa y que son identificados por medio de Inteligencia de Negocios (o Business Intelligence, BI), área de marketing, área de finanzas (principalmente, factores de desempeño económico), como también desde nuevos estándares y/o normativas legales. De esta manera, los factores externos influyen directamente en los internos.

Por lo tanto, luego de haber definido los términos “Proceso” y “Optimizar”, podemos decir que la **Optimización de Procesos** es mejorar y aumentar la productividad de un proceso donde se ven involucrados diversos factores entre los principales tenemos: el recurso humano, la tecnología, la inversión de capital y las reglamentaciones gubernamentales. La optimización de un proceso debe considerar los factores internos y externos de una organización para luego llevarla a cabo. Para optimizar procesos dentro de una organización es recomendable seguir las siguientes pautas:

- Al emplear el término "*Optimización*" se debe dejar en claro las limitaciones de encontrar *el mejor proceso* y que en la práctica sólo encuentra *el que mejor se*

ajuste a la realidad de cada empresa que se ve afectada por factores internos y externos.

- Identificar el *core process* que se quiere optimizar. El *core process* es aquel identificado a partir de la estrategia de negocios de la empresa.
- Identificar los factores internos y externos que afectan la decisión de optimizar un proceso, con el dueño del proceso y dueños de tareas y áreas específicas dentro de la empresa. No olvidar que muchos procesos son transversales a la organización.
- Identificar la variable que quiere "mejorar" dentro de un proceso: tiempo, costo o calidad.
- Aplicar reingeniería, buenas prácticas o rediseño del proceso.
- Simular el nuevo proceso iterativamente hasta encontrar el que mejor se ajuste a nuestros requerimientos.
- Definir medidas de rendimiento del nuevo proceso (KPIs) y monitoréelos.
- Gestionar el cambio del proceso con el dueño del proceso y áreas transversales afectadas.
- Gestionar el conocimiento generado y actualizado en la organización a partir de los cambios realizados al proceso optimizado (o mejorado)
- Monitorear el nuevo proceso e identificar si la ejecución del mismo corresponde al definido y publicado a las partes.

2.2 Proceso de Presupuestación

Según [Cont01], el proceso de presupuestación tiende a reflejar de una forma cuantitativa, a través de los presupuestos, los objetivos fijados por la empresa a corto plazo, mediante el establecimiento de los oportunos programas, sin perder la perspectiva del largo plazo, puesto que ésta condicionará los planes que permitirán la consecución del fin último al que va orientado la gestión de la empresa.

Los presupuestos sirven de medio de comunicación de los planes de toda la organización, proporcionando las bases que permitirán evaluar la actuación de los distintos segmentos, o áreas de actividad de la empresa y de la gerencia.

El proceso culmina con el control presupuestario, mediante el cual se evalúa el resultado de las acciones emprendidas permitiendo, a su vez, establecer un proceso de ajuste que posibilite la fijación de nuevos objetivos.

Un proceso de presupuestación eficaz depende de muchos factores, sin embargo cabe destacar dos que pueden tener la consideración de "requisitos imprescindibles"; así, por un lado, es necesario que la empresa tenga configurada una estructura organizativa clara y coherente, a través de la que se vertebrará todo el proceso de asignación y delimitación de responsabilidades. Un programa de presupuestación será más eficaz en tanto en cuanto se puedan asignar adecuadamente las responsabilidades, para lo cual, necesariamente, tendrá que contar con una estructura organizativa perfectamente definida.

El otro requisito viene determinado por la repercusión que, sobre el proceso de presupuestación, tiene la conducta del potencial humano que interviene en el mismo; esto es, el papel que desempeñan dentro del proceso de planificación y de presupuestación los factores de motivación y de comportamiento. La presupuestación, además de representar un instrumento fundamental de optimización de la gestión a corto plazo, constituye una herramienta eficaz de participación del personal en la determinación de objetivos, y en la formalización de compromisos con el fin de fijar responsabilidades para su ejecución. Esta participación sirve de motivación a los individuos que ejercen una influencia personal, confiriéndoles un poder decisorio en sus respectivas áreas de responsabilidad.

El proceso de planificación presupuestaria de la empresa varía mucho dependiendo del tipo de organización de que se trate, sin embargo, con carácter general, se puede afirmar que consiste en un proceso secuencial integrado por las siguientes etapas:

- *Definición y transmisión de las directrices generales a los responsables de la preparación de los presupuestos*: La dirección general, o la dirección estratégica, es la responsable de transmitir a cada área de actividad las instrucciones generales, para que éstas puedan diseñar sus planes, programas, y presupuestos; ello es debido a que las directrices fijadas a cada área de

responsabilidad, o área de actividad, dependen de la planificación estratégica y de las políticas generales de la empresa fijadas a largo plazo.

- *Elaboración de planes, programas y presupuestos:* A partir de las directrices recibidas, y ya aceptadas, cada responsable elaborará el presupuesto considerando las distintas acciones que deben emprender para poder cumplir los objetivos marcados. Sin embargo, conviene que al preparar los planes correspondientes a cada área de actividad, se planteen distintas alternativas que contemplen las posibles variaciones que puedan producirse en el comportamiento del entorno, o de las variables que vayan a configurar dichos planes.

- *Negociación de los presupuestos*: La negociación es un proceso que va de abajo hacia arriba, en donde, a través de fases iterativas sucesivas, cada uno de los niveles jerárquicos consolida los distintos planes, programas y presupuestos aceptados en los niveles anteriores.

- *Coordinación de los presupuestos*: A través de este proceso se comprueba la coherencia de cada uno de los planes y programas, con el fin de introducir, si fuera necesario, las modificaciones necesarias y así alcanzar el adecuado equilibrio entre las distintas áreas.

- *Aprobación de los presupuestos*: La aprobación, por parte de la dirección general, de las previsiones que han ido realizando los distintos responsables supone evaluar los objetivos que pretende alcanzar la entidad a corto plazo, así como los resultados previstos en base de la actividad que se va a desarrollar.

- *Seguimiento y actualización de los presupuestos*: Una vez aprobado el presupuesto es necesario llevar a cabo un seguimiento o un control de la evolución de cada una de las variables que lo han configurado y proceder a compararlo con las previsiones. Este seguimiento permitirá corregir las situaciones y actuaciones desfavorables, y fijar las nuevas previsiones que pudieran derivarse del nuevo contexto.

2.2.3 Presupuesto

El presupuesto puede ser definido como: un plan integrado y coordinado que se expresa en términos financieros, respecto de las operaciones y recursos que forman parte de una empresa, para un período determinado, con el fin de lograr los objetivos fijados por la alta gerencia.

Según [Cont01], los objetivos que puede perseguir cualquier organización al implantar un presupuesto pueden concretarse en los siguientes aspectos:

- Apoya en las tareas de planificación de las operaciones anuales.
- Permite concretar y cuantificar los objetivos de la alta gerencia, para cada una de las divisiones operativas.
- Motiva a los responsables de los planes definidos en el presupuesto.
- Posibilita evaluar el grado de consecución de los objetivos y planes marcados.
- Permite evaluar las actuaciones del personal directivo, sobre todo cuando lleva aparejado un sistema de compensaciones.
- Potencia las tareas de formación y de desarrollo del personal.

En cuanto a las características generales que conlleva la elaboración de un presupuesto se pueden destacar, resumidamente, las siguientes:

- Pronosticabilidad
- Economicidad
- Susceptibles de revisión
- Flexibilidad
- Fiabilidad
- Participación
- Oportunidad

2.3 Minería de Datos

Dentro de las múltiples áreas que se agrupan en torno a la Gestión del Conocimiento, aparece la Minería de Datos como una de las disciplinas que mas esta influyendo en nuestros días dentro del ámbito del análisis de datos.

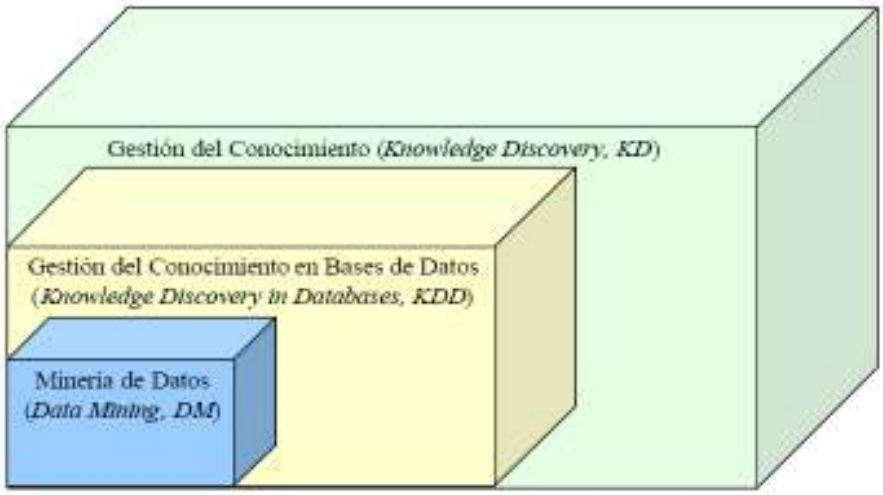


Figura 2.1 La minería de datos frente al KD y KDD. Fuente: [Martinez03]

2.3.1 Definición

El nombre de Minería de Datos (Data Mining en inglés), deriva de la analogía usada para describir la búsqueda de información valiosa de negocios en grandes repositorios de datos, con la búsqueda de vetas de metales preciosos dentro de una determinada área geográfica.

La definición de Minería de Datos puede variar entre unos autores y otros. Dentro del campo sistémico, la mayoría adopta este término para referirse al ***proceso que involucra el empleo de ciertas técnicas y herramientas con la finalidad de extraer información útil de una base de datos.*** Dentro de estas técnicas podemos encontrar todos aquellos métodos y/o algoritmos matemáticos y software para el análisis inteligente de los datos, además de la búsqueda de patrones o tendencias en los mismos, aplicados de forma iterativa e interactiva.

A continuación se presenta una serie de definiciones tomadas de diferentes fuentes, y con las cuales se busca ilustrar de una manera más concreta en qué consiste la Minería de Datos (Data Mining):

- "Data Mining es el proceso analítico diseñado para explorar grandes cantidades de datos (típicamente relacionados con el mercado o los negocios) con el fin de investigar patrones consistentes y/o relaciones sistemáticas entre variables y, a continuación, validar los resultados aplicando modelos detectados para nuevos subgrupos de datos." [Stat01]

- "Data Mining es la extracción de información implícita, previamente desconocida y potencialmente útil de una base de datos." [Witten00]

En cambio, en el ámbito del *Knowledge Discovery in Databases* (KDD) o descubrimiento de conocimiento en bases de datos, la Minería de Datos tiene otro significado, donde éste viene a ser un paso del KDD, y es definido por los autores de la siguiente manera:

- "El proceso de extraer patrones o modelos a partir de los datos." [Fayyad96]

- "Data Mining consiste en obtener modelos comprensibles o patrones de una base de datos." [Siebes00]

- "Las actividades de minería de datos constituyen un proceso iterativo, dirigidas al análisis de grandes bases de datos, con el propósito de extraer información y conocimiento que puede ser exacta y potencialmente útil para los trabajadores del conocimiento encargados de la toma de decisiones y resolución de problemas." [Vercellis09]

Como podemos ver, hay autores que definen a la Minería de Datos como parte del proceso KDD y otros, como el proceso completo de extracción de conocimiento.

En el contexto actual de nuestro estudio tomaremos la Minería de Datos como parte del proceso KDD, tal como se observa en la Figura 2.1, siendo este proceso interactivo e iterativo conteniendo los siguientes pasos: [Riquelme06]

1. ***Comprender el dominio de aplicación:*** este paso incluye el conocimiento relevante previo y las metas de la aplicación.

2. ***Extraer la base de datos objetivo:*** recogida de los datos, evaluar la calidad de los datos y utilizar análisis exploratorio de los datos para familiarizarse con ellos.

3. ***Preparar los datos:*** incluye limpieza, transformación, integración y reducción de datos. Se intenta mejorar la calidad de los datos a la vez que disminuir el tiempo requerido por el algoritmo de aprendizaje aplicado posteriormente.

4. ***Minería de datos:*** como se ha señalado anteriormente, este es la fase fundamental del proceso. Está constituido por una o más de las siguientes

funciones, clasificación, regresión, clustering, resumen, recuperación de imágenes, extracción de reglas, etc.

5. ***Interpretación:*** explicar los patrones descubiertos, así como la posibilidad de visualizarlos.

6. ***Utilizar el conocimiento descubierto:*** hacer uso del modelo creado.

2.3.2 Minería de datos, estadística clásica y OLAP

La minería de datos difiere en mucho respecto a la estadística clásica y análisis OLAP. Aquellas diferencias son mostradas en la Figura 2.2, siendo la principal diferencia la orientación activa ofrecida por los modelos de aprendizaje inductivo, en comparación con el carácter pasivo de las técnicas estadísticas y OLAP. De hecho, en los análisis estadísticos que llevan a la toma de decisiones, es necesario formular una hipótesis que luego ha de ser confirmada sobre la base de la evidencia de la muestra.

Table 5.1 Differences between OLAP, statistics and data mining

OLAP	statistics	data mining
extraction of details and aggregate totals from data	verification of hypotheses formulated by analysts	identification of patterns and recurrences in data
information	validation	knowledge
distribution of incomes of home loan applicants	analysis of variance of incomes of home loan applicants	characterization of home loan applicants and prediction of future applicants

Figura 2.2 Diferencias entre OLAP, Estadística clásica y Minería de Datos. Fuente: [Vercellis09]

Del mismo modo, en el análisis OLAP los trabajadores del conocimiento expresan una intuición basada en la extracción, elaboración de informes y criterios de visualización. Ambos métodos sólo proporcionan elementos para confirmar o refutar las hipótesis formuladas por el tomador de decisiones, según un análisis de flujo de arriba hacia abajo. Por el contrario, los modelos de aprendizaje que constituyen el núcleo de la minería de datos, son capaces de desarrollar un papel activo en la generación de predicciones e interpretaciones que en la realidad aportan los nuevos conocimientos a disposición de los usuarios. [Vercellis09]

El análisis de flujo para este último caso tiene una estructura de abajo hacia arriba. En particular, cuando se enfrentan con grandes cantidades de datos la utilización de modelos capaces de desempeñar un papel activo se convierte en un

factor crítico de éxito, ya que es difícil para el conocimiento de los trabajadores el formular a priori, una significativa y bien fundada hipótesis.

2.4 Series Temporales

Toda institución, ya sea la familia, la empresa o el gobierno, necesita realizar planes para el futuro si desea sobrevivir o progresar. La planificación racional exige prever los sucesos del futuro que probablemente vayan a ocurrir, estas previsiones se suelen basar en los sucesos ocurridos en el pasado. La técnica estadística utilizada para hacer inferencias sobre el futuro teniendo en cuenta lo ocurrido en el pasado es el análisis de series temporales.

Muchas veces se busca calcular un nuevo valor para una variable en cierto período de tiempo futuro, asociado a períodos de tiempo adyacentes; en este caso, la secuencia de los valores de la variable objetivo, representa una serie de tiempo. Por ejemplo, las ventas semanales de un producto dado observado durante dos años representa una serie de tiempo con 104 observaciones. Los modelos para el anállisis de series temporales investigan datos caracterizados por un patrón en el tiempo y apuntan a la predicción del valor de la variable objetivo para uno o varios períodos futuros. [Vercellis09]

"Cuando hablamos de una secuencia de valores observados a lo largo del tiempo, y por tanto ordenados cronológicamente, la denominamos, en un sentido amplio, serie temporal. Resulta difícil imaginar una rama de la ciencia en la que no aparezcan datos que puedan ser considerados como series temporales." [Mol01]

Otros autores definen una serie temporal (también denominada histórica, cronológica o de tiempo) como un conjunto de datos, correspondientes a un fenómeno económico, ordenados en el tiempo.

Si, conocidos los valores pasados de la serie, no fuera posible predecir con total certeza el próximo valor de la variable, decimos que la serie es no determinista o aleatoria, y lógicamente es de éstas de las que se ocupa el cuerpo de doctrina denominado "análisis de series temporales" y al que vamos a dedicar esta breve introducción.

El análisis estadístico de series temporales se usa hoy día con profusión en muchas otras áreas de la ciencia, fundamentalmente en física, ingeniería y en economía.

Los objetivos del análisis de series temporales son diversos, pudiendo destacar la predicción, el control de un proceso, la simulación de procesos, y la generación de nuevas teorías físicas o biológicas.

Denominamos predicción a la estimación de valores futuros de la variable en función del comportamiento pasado de la serie. Este objetivo se emplea ampliamente en el campo de la ingeniería y de la economía, incluyendo en esta última rama también la sanidad pública y la vigilancia de la salud. Así por ejemplo, la predicción mediante modelos basados en la teoría de series temporales, puede servir para una buena planificación de recursos sanitarios, en función de la demanda que se espera en el futuro, prevista por el modelo. Otro de los campos en los que se aplica la predicción mediante series temporales es el de la meteorología o en la predicción de otros fenómenos naturales.

CAPÍTULO 3: Estado del Arte

El almacenamiento de datos se ha convertido en una tarea rutinaria de los sistemas de información, en este sentido los datos almacenados son un valioso tesoro para las organizaciones los cuales representan la memoria de ellas. Pero tener esta memoria no es suficiente, hay que pasar al proceso inteligente de extraer la información que almacenan estos datos para poder generar escenarios, pronósticos y reportes que apoyen a la toma de decisiones, lo que se convierte en una ventaja competitiva. Este es el objetivo que persigue la Minería de Datos. [Aluja01].

Las empresas suelen generar grandes cantidades de información sobre sus procesos productivos, desempeño operacional, mercados y clientes; pero el éxito de los negocios depende de la habilidad para ver nuevas tendencias o cambios en ellas. Las aplicaciones de Minería de Datos pueden identificar tendencias y comportamientos, no sólo para extraer información, sino también para descubrir las relaciones en bases de datos que pueden identificar comportamientos que no son muy evidentes.

La Minería de Datos reúne ventajas de varias áreas como la Estadística, la Inteligencia Artificial, la Computación Gráfica, las Bases de Datos y el Procesamiento Masivo, principalmente usando como materia prima las bases de datos.

Las herramientas comerciales de Minería de Datos que existen actualmente en el mercado son variadas y excelentes. Las hay orientadas al estudio del web o al análisis de documentos o de clientes de supermercado. Su correcta elección depende de la necesidad de la empresa y de los objetivos a corto y largo plazo que pretenda alcanzar.

3.1. Descubrimiento de Conocimiento en Bases de Datos (KDD)

En los últimos años, ha existido un crecimiento en nuestras capacidades de generar y colectar datos, debido al gran poder de procesamiento de los computadores y a su bajo costo de almacenamiento.

Sin embargo, dentro de estas enormes cantidades de datos existe una gran cantidad de información oculta, de gran importancia estratégica, a la que no se puede acceder por las técnicas clásicas de recuperación de la información. El descubrimiento de esta información oculta es posible gracias a la Minería de Datos, que entre varias sofisticadas técnicas aplica la Inteligencia Artificial para encontrar patrones y relaciones dentro de los datos permitiendo la creación de modelos, es decir, representaciones abstractas de la realidad, pero es el descubrimiento del conocimiento (KDD: Knowledge Discovery in Databases) que se encarga de la preparación de los datos y la interpretación de los resultados obtenidos, los cuales dan un significado a estos patrones encontrados.

El valor real de los datos reside en la información que se puede extraer de ellos, información que ayude a tomar decisiones o mejorar nuestra comprensión de lo que nos rodea. Los métodos analíticos son la clave de muchas organizaciones exitosas, incrementando sus ganancias, maximizando la eficiencia operativa, reduciendo costos y mejorando la satisfacción del cliente.

El objetivo fundamental del KDD es encontrar conocimiento útil, válido, relevante y nuevo sobre una actividad mediante algoritmos eficientes. [Vallejos06].

3.2. Problemas abordados por la Minería de Datos

La Minería de Datos resuelve cualquier problema para el que existan datos históricos almacenados.

Las metodologías de minería de datos pueden ser aplicadas a una variedad de ámbitos, desde el marketing y control de procesos de manufactura hasta el estudio de factores de riesgo en diagnósticos médicos, desde la evaluación de la eficacia de nuevos fármacos hasta la detección de fraudes. [Vercellis09].

En [Vallejos06], se indica la siguiente lista ilustrativa de lo abordado por la Minería de datos:

- **Búsqueda de lo inesperado por descripción de la realidad multivariante.** Un principio clásico de la Estadística, el principio de la parsimonia, ya no es ahora válido (si bien siempre serán preferibles los modelos simples). Para describir un fenómeno cuantas más variables tengamos mejor, más ricas, más globales y más coherentes serán las descripciones y más fácil será detectar lo inesperado, esto es, aquello que no habíamos previsto y que resulta valioso para entender mejor el comportamiento de algún grupo de individuos, lo cual se ve favorecido por el hecho de trabajar con muestras grandes. Las muestras aleatorias son suficientes para describir la regularidad estadística global, pero no para detectar comportamientos particulares de subgrupos.

- **Búsqueda de asociaciones.** Un cierto suceso, ¿está asociado a otro suceso?, ¿podemos inferir que determinados sucesos ocurren simultáneamente más de lo que sería esperable si fuesen independientes?, ¿es posible sugerir un producto, sabiendo que otro ha sido adquirido?

- **Definición de tipologías.** Existe una inmensa cantidad de consumidores, pero los tipos de consumidores distintos son un número mucho más pequeño. Detectar estos tipos distintos, su perfil de compra y proyectarlos sobre toda la población, es una operación imprescindible a la hora de programar una política de marketing. Por otro lado, las tipologías no tienen que ser necesariamente de consumo, pueden ser de opiniones, valores, condiciones de vida, etc.

- **Detección de ciclos temporales.** Todo consumidor sigue un ciclo de necesidades que ocasionan actos de compra distintos a lo largo de su vida. Detectar los diferentes ciclos y la fase donde se sitúa cada consumidor ayudará a crear complicidades y adecuar la oferta de productos a las necesidades y crear fidelización.

- **Predicción.** A menudo deberemos efectuar predicciones: ¿cuál es la probabilidad de baja de un cliente?, ¿cuál es el precio de una vivienda concreta?, ¿lloverá mañana? Estas y muchas más son preguntas que deberemos responder, para ello construiremos un modelo a partir de los datos históricos. Si la variable de respuesta es continua (la rentabilidad de un cliente) diremos que se trata de un problema de regresión, mientras que si la variable de respuesta es categórica (p. e. la compra o no de un producto) diremos que se trata de un problema de clasificación.

Los problemas abordados por la minería de datos se ubican dentro algunos de los siguientes ámbitos:

- **Predicción de ventas**: La minería de datos ayuda en este ámbito proporcionando conocimiento valioso acerca de la predicción del volumen de ventas basándose en información histórica. Esto también es válido para la proyección de costos, el cual por ejemplo, puede basarse en el volumen de materia prima a utilizarse en una producción futura, cuando los indicadores de consumo son complejos y variables.

- **Marketing relacional**: Las aplicaciones de minería de datos en el campo del marketing relacional han contribuido significativamente al aumento de la popularidad de esta tecnología. Algunas aplicaciones relevantes del marketing relacional son:
  - Identificación de segmentos de cliente que tienen más probabilidades de responder a ciertas campañas de marketing.
  - Identificación de segmentos de clientes objetivo para la retención de campañas.
  - Predicción de ratios de respuesta positiva a campañas de marketing.
  - Interpretación y comprensión del comportamiento de compra de los clientes.
  - Análisis de los productos comprados por los clientes de forma conjunta, conocido como análisis de la cesta de mercado.

- **Detección de fraudes**: La detección de fraudes es otro ámbito de aplicación de la minería de datos. El fraude puede afectar a diferentes industrias, como la telefonía, los seguros y la banca (uso ilegal de tarjetas de crédito, transacciones monetarias ilícitas).

- **Evaluación de riesgos:** El propósito del análisis de riesgos consiste en estimar el riesgo relacionado con futuras decisiones. Por ejemplo, usando las observaciones pasadas disponibles, un Banco puede desarrollar un modelo predictivo para establecer si es oportuno conceder un préstamo de dinero o un préstamo hipotecario, basado en las características del solicitante.

- **Diagnósticos médicos:** Los modelos de aprendizaje son una herramienta invaluable en el campo de la medicina para la detección precoz de las enfermedades mediante métodos de ensayo clínico. El análisis de imágenes

con fines de diagnóstico es otro campo de investigación que actualmente está en expansión.

3.3. Tareas en Minería de Datos

Cada tarea dentro de la Minería de Datos puede considerarse como un problema diferente a ser resuelto por un algoritmo. Cada tarea tiene sus propios requisitos, y retorna información posiblemente diferente en cada caso.

Las tareas corresponden al modelo (Predictivo o Descriptivo) que pertenezcan, como indica [Cardona05] las más comúnmente utilizadas en los trabajos de Minería de Datos son:

3.3.1 Tareas Predictivas

Se basa en la estimación de valores futuros o desconocidos de variables de interés (variable objetivo) a partir de otras variables independientes (predictivas). Las principales tareas predictivas son: la clasificación y la regresión.

Los algoritmos de clasificación predicen una o más variables discretas, basándose en otros atributos del conjunto de datos, un ejemplo es el algoritmo de árboles de decisión.

Los algoritmos de regresión predicen una o más variables continuas, como las pérdidas o los beneficios, basándose en otros atributos del conjunto de datos, un ejemplo es el algoritmo de series temporales.

El objetivo de la predicción consiste en anticipar el valor que una variable no regular asumirá en el futuro o para estimar la probabilidad de eventos futuros. Por ejemplo, un proveedor de telefonía móvil puede desarrollar un análisis de minería de datos para la estimación de la probabilidad de sus clientes de pasarse a su competidor. En un contexto diferente, una empresa de venta retail podría predecir las ventas de un determinado producto durante las semanas posteriores. [Vercellis09]

En realidad, la mayoría de las técnicas de minería de datos derivan sus predicciones del valor de un conjunto de variables asociadas con las variables de las entidades en una base de datos. Por ejemplo, un modelo de minería de datos podría

indicar la probabilidad en el futuro de retener a un cliente, esto depende de ciertas características, tales como la edad, duración del contrato y el porcentaje de llamadas a un abonado de otros proveedores de telefonía. Sin embargo, existen los modelos de series de tiempo que hacen predicciones basadas en los valores pasados de la variable de interés, en el cual centraremos nuestro estudio.

Muchas veces, un modelo desarrollado con el propósito de predicción también puede convertirse eficaz para la interpretación. En el caso de los árboles de clasificación, los modelos generados para fines de predicción también pueden resultar útiles en la identificación de fenómenos de motivaciones periódicas.

3.3.2 Tareas Descriptivas

Identificación de patrones en los datos que los explican o resumen. Las principales tareas descriptivas: son el agrupamiento o segmentación, y la asociación.

Los algoritmos de segmentación dividen los datos en grupos o clústeres de elementos que tienen propiedades similares, un ejemplo es el algoritmo de clústeres.

Los algoritmos de asociación buscan correlaciones entre diferentes atributos de un conjunto de datos. La aplicación más común de esta clase de algoritmo es la creación de reglas de asociación, que pueden utilizarse en un análisis de la cesta de compra.

El propósito de la interpretación es identificar patrones regulares en los datos y de expresarse a través de normas y criterios que puedan ser fácilmente entendidos por los expertos en el dominio de la aplicación. Las reglas generadas deben ser originales y no triviales, a fin de aumentar realmente el nivel de conocimiento y comprensión del sistema de interés. Por ejemplo, para una compañía en la industria retail podría ser ventajoso agrupar a los clientes que vayan a sacar su tarjeta de fidelidad, de acuerdo a su perfil de compra. De este modo, los segmentos generados podrían resultar útiles en la identificación de nuevos nichos de mercado y de la dirección de futuras campañas de marketing. [Vercellis09]

3.4. Técnicas de Minería de Datos

Las técnicas son la particularización de las tareas anteriormente mencionadas. Por medio de un algoritmo es posible resolver varias tareas, pero existen ventajas e inconvenientes que permiten identificar cuál algoritmo es el más apropiado.

Las técnicas de la minería de datos provienen de la Inteligencia Artificial y de la Estadística, estas técnicas no son más que algoritmos más o menos sofisticados que se aplican sobre un conjunto de datos para obtener resultados.

[Cardona05] señala que ”la conveniencia de aplicar un determinado algoritmo depende no sólo del tipo de problema con el que nos estemos enfrentando sino que depende en gran medida del tipo de los datos con los que se está tratando. En este sentido conviene analizar los distintos enfoques y algoritmos que existen en la literatura”.

Existen diferentes algoritmos y variaciones de los mismos, así como restricciones que hacen que la efectividad del algoritmo dependa del dominio de aplicación, por lo que no es posible encontrar un método universal aplicable a todas las situaciones.

Un enriquecimiento de las posibilidades de análisis son los sistemas híbridos, esto es, la combinación de dos o más técnicas para mejorar la eficiencia en la resolución de un problema, como por ejemplo, utilizar un algoritmo genético para inicializar una red neuronal, o bien utilizar un árbol decisión como variable de entrada en una regresión logística.

En [Nadinic08] se muestra un cuadro comparativo entre las principales tareas y técnicas de Minería de Datos, como podemos observar en la Tabla 3.1:

	Predictivas		Descriptivas	
Técnica	Clasificación	Regresión	Agrupamiento	Asociación
Redes Neuronales	✓	✓	✓	
Redes de Kohonen			✓	
Árboles de decisión (ID.3, C4.5)	✓			
Árboles de decisión (CART)	✓	✓		
Regresión lineal		✓		
Clustering (K-means)			✓	
Clustering (Vecinos más próximos K-NN)	✓	✓	✓	
Redes Bayesianas	✓			
Series Temporales		✓		
Previsión Local	✓	✓	✓	
Algoritmos genéticos	✓	✓	✓	✓

Tabla 3.1 Técnicas de Minería de Datos. Fuente: [Nadinic08]

Como se señala en la Tabla 3.1, cualquiera que sea el problema a resolver, no existe una única técnica para solucionarlo. En [Vallejos06], se describen las siguientes técnicas:

3.4.1 Redes Neuronales

Inspiradas en el modelo biológico, son generalizaciones de modelos estadísticos clásicos. Como indica [Nadinic08] se trata de técnicas que aprenden un modelo mediante el entrenamiento de los pesos que conectan un conjunto de nodos o neuronas, la topología de la red y los pesos de las conexiones determinan el patrón aprendido.

Su novedad radica en el aprendizaje secuencial, el hecho de utilizar transformaciones de las variables originales para la predicción y la no linealidad del modelo.

Una de las principales características de las redes neuronales, es que son capaces de trabajar con datos incompletos e incluso paradójicos, que dependiendo del problema puede resultar una ventaja o un inconveniente. Además posee dos formas de aprendizaje: supervisado y no supervisado.

Permite aprender en contextos difíciles, sin precisar la formulación de un modelo concreto. Su principal inconveniente es que para el usuario son una caja negra.

Ejemplos de redes neuronales artificiales son:

- Perceptón simple. Se refiere a la neurona artificial y unidad básica de inferencia en forma de discriminador lineal, que constituye este modelo de red neuronal artificial, esto debido a que el perceptrón puede usarse como neurona dentro de un perceptrón más grande u otro tipo de red neuronal artificial.

- Perceptrón multicapa. Es una red neuronal artificial formada por múltiples capas, esto le permite resolver problemas que no son linealmente separables, lo cual es la principal limitación del perceptrón simple. Es usado para resolver problemas de asociación de patrones, segmentación de imágenes, compresión de datos, etc.

- Mapa Autoorganizado o redes de Kohonen. Son un tipo de red neuronal no supervisada, competitiva, distribuida de forma regular en una rejilla de dos dimensiones, cuyo fin es descubrir la estructura subyacente de los datos introducidos en ella. A lo largo del entrenamiento de la red, los vectores de datos son introducidos en cada neurona y se comparan con el vector de peso característico de cada neurona. La neurona que presenta menor diferencia entre su vector de peso y el vector de datos es la neurona ganadora y ella y sus vecinas verán modificados sus vectores de pesos.

3.4.2 Regresión Lineal

Es la técnica más utilizada para formar relaciones entre datos. Rápida y eficaz pero insuficiente en espacios multidimensionales donde puedan relacionarse más de 2 variables.

Son modelos que permiten tratar diferentes tipos de variables de respuesta, por ejemplo la preferencia entre productos concurrentes en el mercado. Al mismo tiempo, los modelos estadísticos se enriquecen cada vez más y se hacen más flexibles y adaptativos, permitiendo abordar problemas cada vez más complejos: GAM, Projection Pursuit, PLS, MARS.

3.4.3 Árboles de Decisión

Este modelo de predicción se encuentra dentro de una metodología de aprendizaje supervisado.

Su representación es en forma de árbol en donde cada nodo es una decisión, los cuales a su vez generan reglas para la clasificación de un conjunto de datos. Permiten obtener de forma visual las reglas de decisión bajo las cuales operan los consumidores, a partir de datos históricos almacenados.

Su principal ventaja es la facilidad de interpretación. Los árboles de decisión son fáciles de usar, admiten atributos discretos y continuos, tratan bien los atributos no significativos y los valores faltantes.

Ejemplos de árboles de decisión son:

- Algoritmo ID3. Este algoritmo es utilizado dentro del ámbito de la inteligencia artificial, fue introducido por Quinlan en el año 1986. Su uso se engloba en la búsqueda de hipótesis o reglas en él, dado un conjunto de ejemplos. El conjunto de ejemplos deberá estar conformado por una serie de tuplas de valores, cada uno de ellos denominados atributos, en el que uno de ellos es el objetivo, el cual es de tipo binario (positivo o negativo, si o no, válido o inválido). De esta forma el algoritmo trata de obtener las hipótesis que clasifiquen ante nuevas instancias, si dicho ejemplo va a ser positivo o negativo. ID3 realiza esta labor mediante la construcción de un árbol de decisión.

- Algoritmo C4.5. Quinlan (1993) propone una mejora del algoritmo ID3, al que denomina C4.5. Este algoritmo se basa en la utilización del criterio ratio de ganancia, de esta manera se evita que las variables con mayor número de posibles valores salgan beneficiadas en la selección. Además, el algoritmo

C4.5 incorpora una poda de árbol de clasificación una vez que éste ha sido inducido, la poda está basada enla aplicación de un test de hipótesis que trata de responder a la pregunta de si merece la pensa expandir o no una determinada rama.

- Algoritmo CART. Este algoritmo no requiere partición a priori, las condiciones en el árbol están basada en umbrales que se calculan dinámicamente; por ello, a lo largo del árbol se puede usar una característica varias veces con diferentes umbrales. CART, permite introducir nuevas características, renstringidas a combinaciones lineales de las existentes.

3.4.4 Agrupamiento o Clustering

Son técnicas que parten de una medida de proximidad entre individuos y a partir de ahí, buscar los grupos de individuos más parecidos entre sí, según una serie de variables mesuradas.

Su utilización ha proporcionado significativos resultados en lo que respecta a los clasificadores o reconocedores de patrones, como en el modelado de sistemas.

Este método, debido a su naturaleza flexible, se puede combinar fácilmente con otro tipo de técnica de minería de datos, dando como resultado un sistema híbrido.

Un problema relacionado con el análisis de cluster es la selección de factores en tareas de clasificación, debido a que no todas las variables tienen la misma importancia a la hora de agrupar los objetos. Otro problema de gran importancia es la fusión de conocimiento, ya que existen múltiples fuentes de información sobre un mismo tema, los cuales no utilizan una categorización homogénea de los objetos. Para poder solucionar estos inconvenientes es necesario fusionar la información a la hora de recopilar, comparar o resumir los datos.

Ejemplos de clustering son:

- Algoritmo K-means. Se utiliza para encontrar los *k* puntos más densos en un conjunto arbitrario de puntos. El algoritmo de k-means clustering es el referente principal entre los diversos métodos para seleccionar grupos representativos entre los datos.

- Algoritmo de vecinos más próximos (Algoritmo K-NN). Es un algoritmo de clasificación supervisada que sirve para estimar la función de densidad de las predictoras por cada clase. Este es un método de clasificación no paramétrico, que estima el valor de la función densidad de probabilidad o directamente la probabilidad a posteriori de que un elemento pertenezca a una clase a partir de la información proporcionada por el conjuto de prototipos. En el proceso de aprendizaje no se hace ninguna suposición acerca de la distribución de las variables predictorias. Es un método de clasificación de objetos o elementos basado en un entrenamiento mediante ejemplos cercanos en el espacio de los elementos.

3.4.5 Series Temporales

Una serie temporal es una sucesión ordenada en el tiempo de valores de una variable. Aunque el tiempo es una variable continua, en la práctica se utilizan las mediciones discretas correspondientes a períodos equidistantes en el tiempo.

A partir de la serie de comportamiento histórica, permite modelizar las componentes básicas de la serie, tendencia, ciclo y estacionalidad y así poder hacer predicciones para el futuro, tales como cifra de ventas, previsión de consumo de un producto o servicio, etc.

3.4.6 Redes Bayesianas

Consiste en representar todos los posibles sucesos en que estamos interesados mediante un grado de probabilidades condicionales de transición entre sucesos. Puede codificarse a partir del conocimiento de un experto o puede ser inferido a partir de los datos. Permite establecer relaciones causales y efectuar predicciones haciendo uso del teorema de Bayes.

3.4.7 Previsión local

La idea de base es que individuos parecidos tendrán comportamientos similares respecto de una cierta variable de respuesta. La técnica consiste en situar los individuos en un espacio euclídeo y hacer predicciones de su comportamiento a partir del comportamiento observado en sus vecinos.

3.4.8 Algoritmos genéticos

Los algoritmos genéticos imitan la evolución de las especies mediante la mutación, reproducción y selección, como también proporcionan programas y optimizaciones que pueden ser usadas en la construcción y entrenamiento de otras estructuras como es el caso de las redes neuronales. Además los algoritmos genéticos son inspirados en el principio de la supervivencia de los más aptos. Es una técnica muy prometedora. En principio cualquier problema que se plantee, como la optimización de una combinación entre distintas componentes, estando estas componentes sujetas a restricciones, puede resolverse mediante algoritmos genéticos.

3.5. Nuevas Tendencias en Técnicas de Minería de Datos

La tendencia actual en las técnicas de Minería de Datos es la de integrar dos puntos de vista, provenientes de la estadística y de la Inteligencia Artificial, en las soluciones algorítmicas propuestas anteriormente, de forma de aprovechar los puntos fuertes de ambas disciplinas. En consecuencia los algoritmos deberían contemplar las dos siguientes propiedades básicas:

- *Poder de generalización a poblaciones diferentes de la observada.* Lo cual implica implementar técnicas eficientes de validación de resultados, ya sea a partir del conocimiento de la distribución muestral de los estadísticos del modelo o por métodos computacionales como la validación cruzada, etc.

- *Escalabilidad.* Dado el volumen de datos a tratar, el coste de los algoritmos ha de ser todo lo lineal que sea posible respecto de los parámetros que definen el coste, en particular respecto del número de individuos.

Las técnicas de Minería de Datos han evolucionado en los últimos años y se han adaptado a los cambios de la tecnología, como el internet, el análisis de datos ocultos en texto y en símbolos, que a continuación se explicarán.

3.5.1 Web Mining

Es importante el análisis de datos recibidos por internet y «on line», dando lugar al **web mining**, donde las técnicas de data mining se utilizan para optimizar las interacciones a través de la web. ¿Cuáles son las secuencias de páginas más

visitadas?, ¿qué páginas visitan los que compran?, ¿los que compran, vuelven a conectarse?, ¿una vez efectuada una adquisición, qué productos puedo sugerir?, son algunas de las preguntas que los responsables de comercio electrónico de las empresas se están formulando en estos momentos.

3.5.2 Text Mining

También los datos objeto de análisis pueden ser textos, dando lugar al **text mining**. Esto es particularmente útil en el análisis de las encuestas de satisfacción percibida por los usuarios. La utilización de las frases realmente escritas supone un enriquecimiento de los análisis realizados sólo con información numérica. También la utilización del text mining para la síntesis y la presentación de la información encontrada en la web es un campo actual de investigación. Más a largo plazo podrán utilizarse la voz o las imágenes.

3.5.3 Fuzzy Mining

Otra de las nuevas vías de investigación es el **fuzzy mining**, esto es, la utilización de las técnicas de minería de datos con objetos simbólicos, que representen más fidedignamente la incertidumbre que se tiene de los objetos que se estudian.

3.6 Metodologías de desarrollo de proyectos de Minería de Datos

Los proyectos de Minería de Datos tienen por finalidad extraer información útil a partir de grandes volúmenes de datos, los cuales se aplican a todos los sectores y campos. Es así que existen proyectos de este tipo en sectores tan dispares como el comercio electrónico, educativo, banca, industrias y/o exploración petrolífera. La extracción de esta información útil se da dentro un proceso complejo, el cual requiere la aplicación de una metodología estructurada para una óptima utilización de las técnicas y herramientas disponibles.

En este sentido, se presentan las principales metodologías utilizadas en la actualidad por los analistas para la realización de proyectos de Minería de Datos: CRISP-DM y SEMMA. Estas metodologías comparten la misma esencia estructurando el proyecto de Minería de Datos en fases que se encuentran interrelacionadas entre sí, convirtiendo el proceso de Minería de Datos en un proceso iterativo e interactivo

[Gondar01]. La presentación de las diferentes fases y tareas de cada metodología proporciona una idea más amplia respecto a la realización de proyectos de Minería de Datos, que facilitará la adaptación de las metodologías, al desarrollo de los proyectos de Minería de Datos específicos de cada organización. Así mismo, la presentación de las fortalezas y debilidades de cada una de las metodologías hace posible la selección informada de una técnica de desarrollo apropiada para cada caso, lo cual servirá como base para la implementación de nuestra solución en el presente trabajo de investigación.

3.6.1 Metodología Semma

SAS Institute, desarrollador de esta metodología, la define como el proceso de selección, exploración y modelado de grandes cantidades de datos para descubrir patrones de negocio desconocidos. El nombre de esta terminología es el acrónimo correspondiente a las cinco fases básicas del proceso. La Figura 3.1 muestra el flujo del proceso de esta metodología que se da en cinco niveles.

El proceso se inicia con la extracción de la población muestral sobre la que se va a aplicar el análisis. El objetivo de esta fase consiste en seleccionar una muestra representativa del problema en estudio. La representatividad de la muestra es indispensable ya que de no cumplirse invalida todo el modelo y los resultados dejan de ser admisibles. La forma más común de obtener una muestra es la selección al azar, es decir, cada uno de los individuos de una población tiene la misma posibilidad de ser elegido. Este método de muestreo se denomina muestreo aleatorio simple.

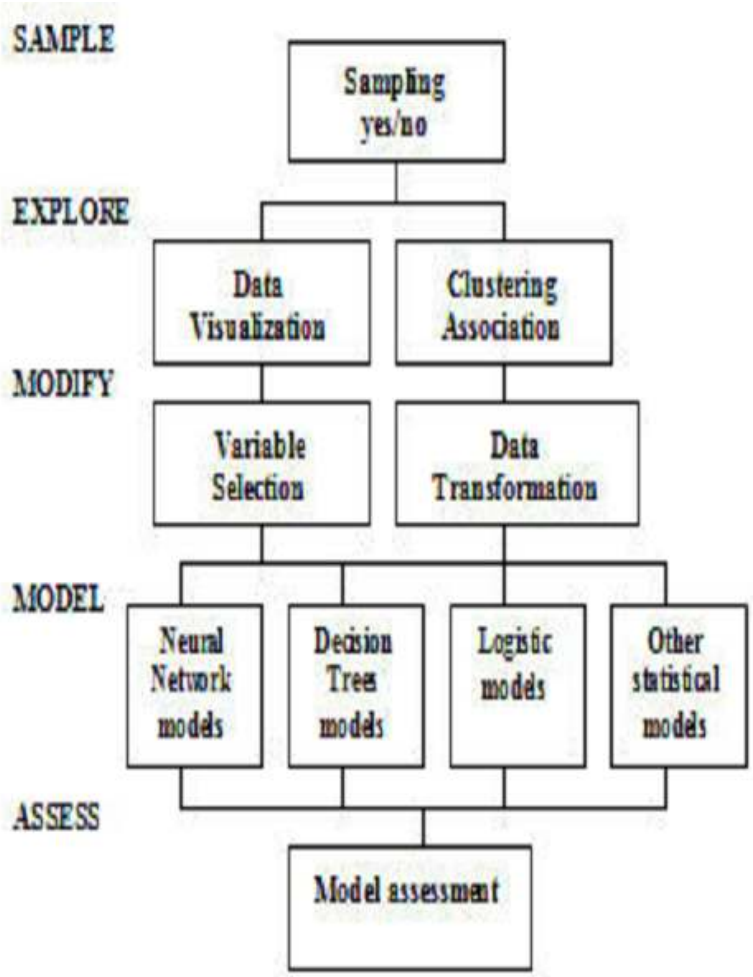


Figura 3.1. Diagrama de Flujo para el diseño SEMMA. Fuente: [Mantignon01]

La metodología SEMMA establece que para cada muestra considerada para el análisis del proceso se debe asociar el nivel de confianza de la muestra. Una vez determinada una muestra o conjunto de muestras representativas de la población en estudio, la metodología SEMMA indica que se debe proceder a una exploración de la información disponible con el fin de simplificar en lo posible el problema con el fin de optimizar la eficiencia del modelo. Para lograr este objetivo se propone la utilización de herramientas de visualización o de técnicas estadísticas que ayuden a poner de manifiesto relaciones entre variables. De esta forma se pretende determinar cuáles son las variables explicativas que van a servir como entradas al modelo.

La tercera fase de la metodología consiste en la manipulación de los datos, en base a la exploración realizada, de forma que se definan y tengan el formato adecuado los datos que serán introducidos en el modelo.

Una vez que se han definido las entradas del modelo, con el formato adecuado para la aplicación de la técnica de modelado, se procede al análisis y modelado de los datos. El objetivo de esta fase consiste en establecer una relación entre las variables explicativas y las variables objeto del estudio, que posibiliten inferir el valor de las mismas con un nivel de confianza determinado. Las técnicas utilizadas para el modelado de los datos incluyen métodos estadísticos tradicionales (tales como análisis discriminante, métodos de agrupamiento, y análisis de regresión), así como técnicas basadas en datos tales como redes neuronales, técnicas adaptativas, lógica fuzzy, árboles de decisión, reglas de asociación y computación evolutiva.

Finalmente, la última fase del proceso consiste en la valoración de los resultados mediante el análisis de bondad del modelo o modelos, contrastado con otros métodos estadísticos o con nuevas poblaciones muestrales.

3.6.2 Metodología CRISP-DM

La metodología CRISP-DM (CRoss Industry Standard Process for Data Mining) desarrollado en 1999 [Chapman99] consta de cuatro niveles de abstracción, organizados de forma jerárquica en tareas que van desde el nivel más general hasta los casos más específicos. Los cuatro niveles son los siguientes:

- **Nivel 1:** Se le denomina fase al asunto o paso dentro del proceso. CRISP-DM consta a su vez de 6 fases:

  - *Comprensión del negocio*: Fase inicial que se enfoca en la comprensión de los objetivos del proyecto y exigencias desde una perspectiva del negocio. En esta fase se define el problema
  - *Comprensión de los datos*: Se recolectan los datos iniciales, familiarización con los datos, se identifican los problemas de calidad de datos, descubrimiento de relaciones iniciales que formaran hipótesis en cuanto a la información oculta.
  - *Preparación de los datos*: Se construye el conjunto de datos final, a partir de los datos brutos iniciales. Las tareas a desarrollar incluyen la selección de tablas, registros y atributos, así como también la transformación y limpieza de datos para las herramientas que se encargarán de modelar.
  - *Modelamiento*: Se eligen y aplican varias técnicas de modelado. Normalmente existen varias técnicas para un mismo problema de

minería de datos. Algunas de ellas tienen requerimientos específicos sobre la forma de datos.

- *Evaluación*: En esta etapa, se tiene un modelo adecuado para realizar un análisis de datos. Antes de proceder al despliegue final del modelo, es importante revisar a fondo los pasos que se han seguido para crearlo, para comparar el modelo correctamente obtenido con los objetivos de negocio
- *Explotación*: Es la obtención del conocimiento a partir del modelo obtenido, este conocimiento ganado tendrá que ser organizado y presentado de un modo en el que el usuario final pueda usarlo. Dependiendo de los requerimientos, la fase de desarrollo puede ser tan simple como la generación de un informe o tan compleja como la realización repetida de un proceso cruzado de minería de datos a través de la organización.

La secuencia entre las fases es dinámica, y es necesario el movimiento hacia delante y hacia atrás entre fases diferentes como se aprecia en la Figura 3.2. Las flechas señalan las más importantes y frecuentes dependencias entre fases. El círculo externo simboliza la naturaleza cíclica de la minería de datos. Esta metodología tiene un modelo cíclico en el que en cada iteración obtiene más cantidad de información y de más calidad para el negocio.

- **Nivel 2:** *Tarea genérica*, cada fase esta formada por tareas genéricas, o sea, la tarea genérica es la descripción de las actividades que se realizan dentro de cada fase. Por ejemplo, la tarea Limpiar los datos es una tarea genérica.

- **Nivel 3:** *Tarea especializada,* la cual describe cómo se pueden llevar a cabo las tareas genéricas en situaciones específicas. Por ejemplo, la tarea Limpiar los datos tiene tareas especializadas, como limpiar valores numéricos, y limpiar valores categóricos.
- **Nivel 4:** *Instancias de proceso*, las cuales son las acciones y resultados de las actividades realizadas dentro de cada fase del proyecto.

A nivel más general, el proceso está organizado en seis fases, estando cada fase a su vez estructurada en varias tareas generales de segundo nivel. Las tareas generales se proyectan a tareas específicas, donde se describen las acciones que deben ser desarrolladas para situaciones específicas. Así, si en el segundo nivel se tiene la tarea general “limpieza de datos”, en el tercer nivel se dicen las tareas que

tienen que desarrollarse para un caso específico, como por ejemplo, “limpieza de datos numéricos”, o “limpieza de datos categóricos”. El cuarto nivel, recoge el conjunto de acciones, decisiones y resultados sobre el proyecto de Minería de Datos específico.

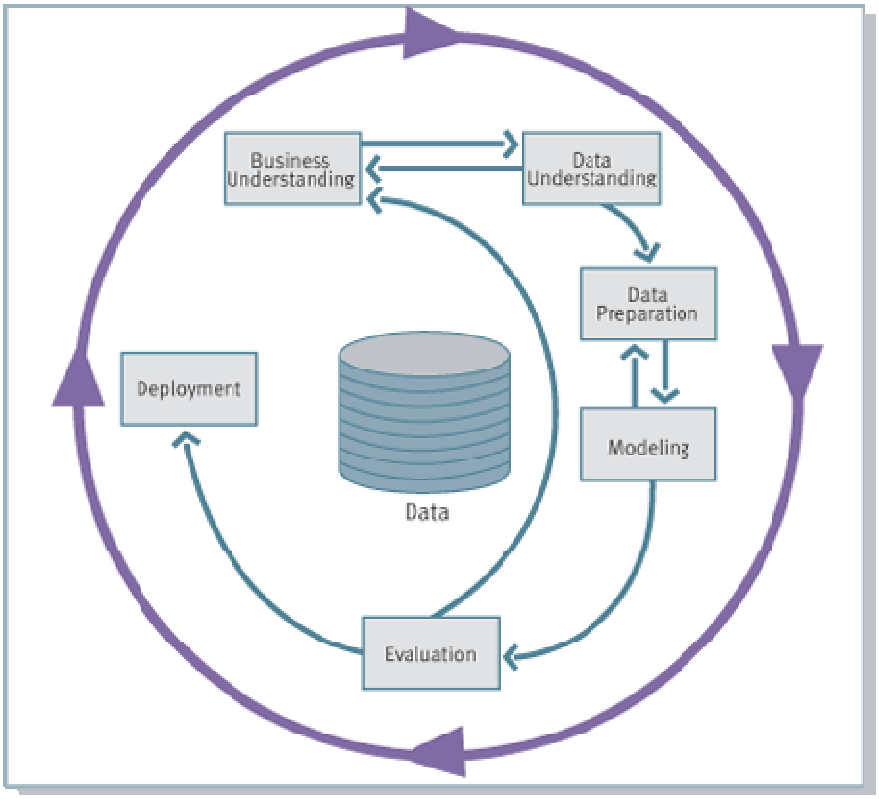


Figura 3.2. Fases del Modelo de Proceso CRISP-DM. Fuente: [Crisp01]

La metodología CRISP-DM proporciona dos documentos distintos como herramienta de ayuda en el desarrollo del proyecto de Minería de Datos: el modelo de referencia y la guía del usuario. El documento del modelo de referencia describe de forma general las fases, tareas generales y salidas de un proyecto de Minería de Datos en general. La guía del usuario proporciona información más detallada sobre la aplicación práctica del modelo de referencia a proyectos de Minería de Datos específicos, proporcionando consejos y listas de comprobación sobre las tareas correspondientes a cada fase.

La metodología CRISP-DM estructura el ciclo de vida de un proyecto de Minería de Datos en seis fases, que interactúan entre ellas de forma iterativa durante el desarrollo del proyecto.

La primera fase análisis del problema, incluye la comprensión de los objetivos y requerimientos del proyecto desde una perspectiva empresarial, con el fin de convertirlos en objetivos técnicos y en una planificación. La segunda fase de análisis

de datos comprende la recolección inicial de datos, en orden a que sea posible establecer un primer contacto con el problema, identificando la calidad de los datos y estableciendo las relaciones más evidentes que permitan establecer las primeras hipótesis. Una vez realizado el análisis de datos, la metodología establece que se proceda a la preparación de los datos, de tal forma que puedan ser tratados por las técnicas de modelado. La preparación de datos incluye las tareas generales de selección de datos a los que se va a aplicar la técnica de modelado (variables y muestras), limpieza de los datos, generación de variables adicionales, integración de diferentes orígenes de datos y cambios de formato. La fase de preparación de los datos, se encuentra muy relacionada con la fase de modelado, puesto que en función de la técnica de modelado que vaya a ser utilizada los datos necesitan ser procesados en diferentes formas. Por lo tanto las fases de preparación y modelado interactúan de forma sistemática. En la fase de modelado se seleccionan las técnicas de modelado más apropiadas para el proyecto de Minería de Datos específico. Las técnicas a utilizar en esta fase se seleccionan en función de los siguientes criterios:

- Ser apropiada al problema.
- Disponer de datos adecuados.
- Cumplir los requerimientos del problema.
- Tiempo necesario para obtener un modelo.
- Conocimiento de la técnica.

Antes de proceder al modelado de los datos se debe de establecer un diseño del método de evaluación de los modelos, que permita establecer el grado de bondad de los modelos. Una vez realizadas estas tareas genéricas se procede a la generación y evaluación del modelo. Los parámetros utilizados en la generación del modelo dependen de las características de los datos.

En la fase de evaluación, se evalúa el modelo, no desde el punto de vista de los datos, sino del cumplimiento de los criterios de éxito del problema. Se debe revisar el proceso seguido, teniendo en cuenta los resultados obtenidos, para poder repetir algún paso en el que, a la vista del desarrollo posterior del proceso, se hayan podido cometer errores. Si el modelo generado es válido en función de los criterios de éxito establecidos en la primera fase, se procede a la explotación del modelo. Normalmente los proyectos de Minería de Datos no terminan en la implantación del modelo, sino que se deben documentar y presentar los resultados de manera comprensible en orden a lograr un incremento del conocimiento. Además en la fase de explotación se debe de asegurar el mantenimiento de la aplicación y la posible difusión de los resultados.

3.7 Modelos usados en Minería para presupuestar

En las operaciones mineras se usan varios modelos para elaborar un presupuesto, los más conocidos en la Gran Minería son el Modelo Financiero Xeras de Runge y el Modelo Hyperion Planning and Budgeting de Oracle, los cuales vamos a detallar a continuación.

3.7.1 Modelo Financiero XERAS

Las operaciones mineras, con la finalidad de optimizar el proceso de presupuestación debido al nuevo plan de cuentas contables basado en Actividades (Costeo ABC) vigente a partir de Enero 2010 en Perú, han adoptado un nuevo modelo de costos para tajo abierto "XERAS 7.7 Financial Modelling", este modelo ha sido desarrollado por Runge Limited, Brisbane, Australia, y ha sido personalizado a las especificaciones de cada una de las 26 minas productivas con las que cuenta Barrick Gold Corporation en todo el mundo (Sudamérica: Perú, Chile y Argentina; Centroamérica: República Dominicana; Norteamérica: Estados Unidos y Canadá; Africa: Tanzania; y Australia Pacífico).

Xeras es un sistema de costeo basado en actividades diseñado para apoyar en la toma de decisiones estratégicas, la planificación a largo plazo y el presupuesto anual.

Xeras ha sido construido para manejar cálculos de equipos mediante variables que permiten auditar las fórmulas y hacer un seguimiento de los datos de los reportes financieros volviendo a la fuente de datos de donde sale estos cálculos.

Xeras es una base de datos estructurada basada en hojas de cálculo y otras tablas de datos; las hojas de cálculo son totalmente definibles por el usuario y se ingresan al sistema mediante plantillas definidas también por el usuario, estos datos provenientes de las hojas de cálculo sirven de indicadores claves para el cálculo de los principales consumibles, como el diesel, explosivos, accesorios de voladura y aceros de perforación, todo esto para el caso del proceso de presupuestación en Operaciones Mina. El uso de plantillas mejoran la eficiencia en la construcción y capacidad de auditoría sobre los sistemas de hoja de cálculo convencional. Los enlaces visuales de hojas de cálculo en un modelo XERAS, así como el uso de las estructuras de los

modelos permite un mayor detalle en el desarrollo, utilización y mantención de datos de manera confiable. En la Figura 3.3 se muestra la interface gráfica del Xeras.

"The data contained in XERAS is accurate, reliable and relevant. In just a couple of minutes a new mine plan can be evaluated, the forecasted costs by cost center can be dumped into an SAP upload report, and the new forecast brought into SAP … for BHP Billiton's Ekati Mine, the XERAS system is seen as the way to go for future budget seasons." [Runge01].

Como se indica en el párrafo anterior, al ingresar datos exactos, confiables y relevantes en XERAS, éste provee de cálculos auditables que son confiables. En unos pocos minutos se puede correr el modelo completo en base a nuevos indicadores, cambios de precios de consumibles mayores (major cost drivers), requerimientos de equipos o nuevos planes de minado; se pueden trabajar con varios escenarios de presupuesto apoyando a la elección del presupuesto más rentable de acuerdo a lo que busca la Compañía.

El modelo contiene una serie de elementos tales como un calendario para definir períodos de tiempo, contiene un anexo para almacenar la data del escenario de producción, contiene estructuras para las tareas de modelización, hojas de recursos que contienen bases de datos sobre los diversos tipos de equipos mineros, tipos de jornadas laborales, entre otros.

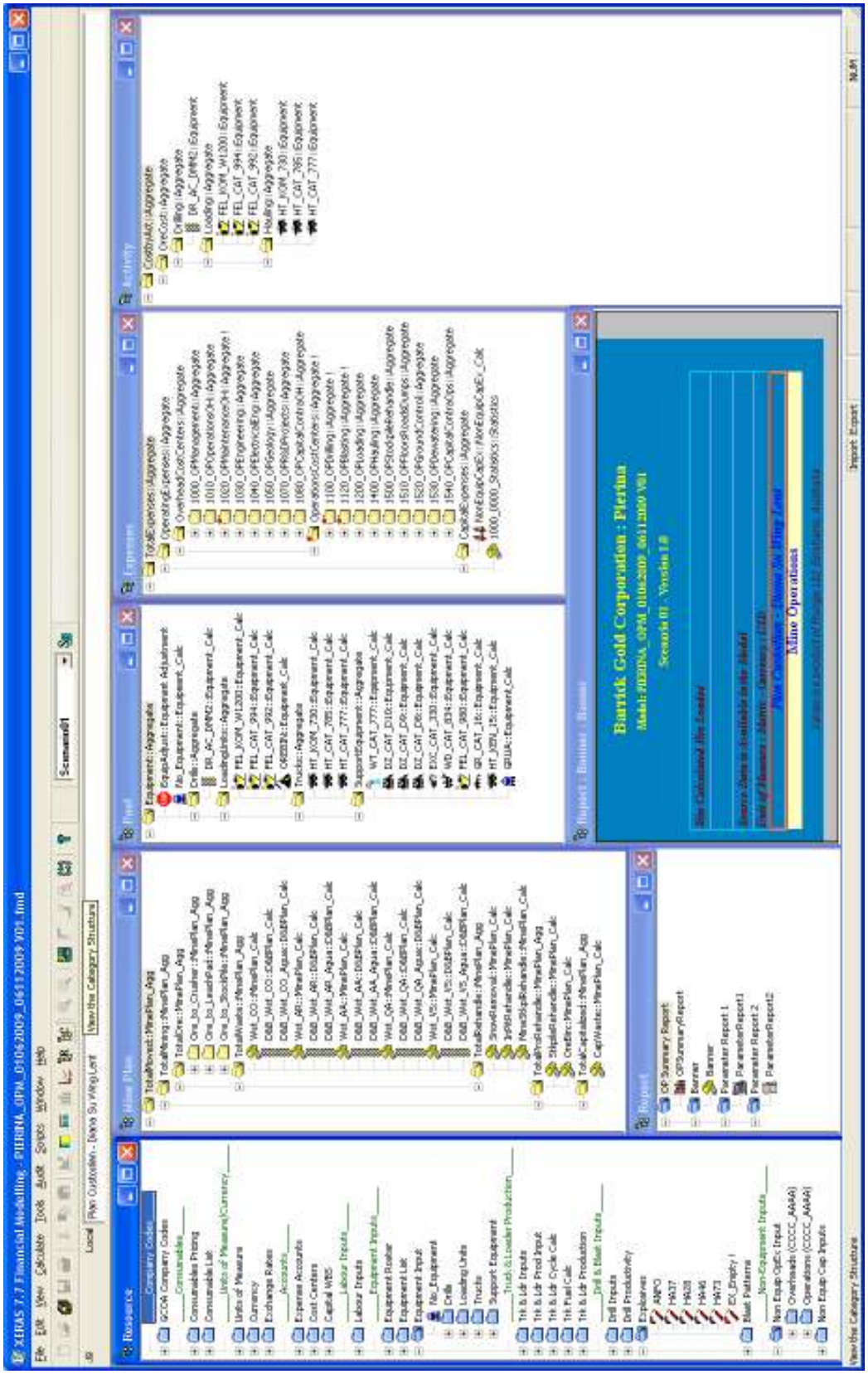


Figura 3.3 Modelo Xeras. Fuente: Minera Barrick Misquichilca. 2009.

3.7.2 Oracle Hyperion Plannning

Hyperion Planning-System 9 es una solución centralizada de elaboración de planificaciones, presupuestos y previsiones basada en Excel y en web, que integra procesos de planificación financiera y operativa. La planificación proporciona una visión profunda de las operaciones de negocio y su impacto derivado sobre las finanzas, mediante una integración estrecha de los modelos de planificación financiera y operativa. La planificación le permite satisfacer las necesidades inmediatas de planificación financiera mientras habilita una plataforma para la futura expansión interfuncional y la integración de procesos automatizada.

Oracle ha adquirido Hyperion hace un par de años con la finalidad de reorientar significativamente su suite de productos de Inteligencia de Negocios y así poder ofrecer a las organizaciones la posibilidad de poder trabajar con data proveniente y no proveniente de Oracle. [Oracle01]. Es por ello que, poco después de esta adquisición, Oracle introdujo una nueva familia de productos llamada Oracle Business Intelligence Enterprise Edition Plus; en Julio del 2008 sale la nueva versión de Oracle Enterprise Performance Management System, que incluye mayores innovaciones y capacidades para mejorar las perspectivas del negocio y la toma de decisiones.

Hyperion Planning es alimentado con data proveniente de XERAS a través de exportaciones del mismo a hojas de cálculo en Excel, no realiza cálculo de variables para poder evaluar escenarios de planes de minado o de presupuestos, por lo que queda limitado a sólo recibir data y mostrarla en reportes.

Hyperion también se alimenta manualmente o a través de formularios importados de hojas de cálculo de data real de cantidades y precios de consumibles, como se puede observar en la Figura 3.4.

Según [Oracle02], las principales ventajas del Hyperion Planning son:

- *Precisión garantizada*: Validación de las previsiones con la mejor analítica integrada.

- *Reduzca el tiempo de elaboración de presupuestos:* Acorte la duración de los ciclos en semanas o meses.

- *Alineación de la organización:* Combinación de la planificación financiera y operativa en un solo sistema.

- *Respuesta veloz a las necesidades financieras:* Cumplimiento inmediato de los requisitos financieros mientras se habilitan procesos de elaboración de presupuestos específicos para las operaciones.

- *Maximización de las capacidades de modelado:* Capacidades de modelado de usuario potentes y avanzadas mediante la integración casi directa con Microsoft Excel.

- *Aceptación masiva de usuarios:* Captación de comunidades de usuarios más amplias con una interfaz web simplificada.

- *Reduzca el tiempo de Implementación:* Con módulos de planificación funcional integrados rápidos de implantar.

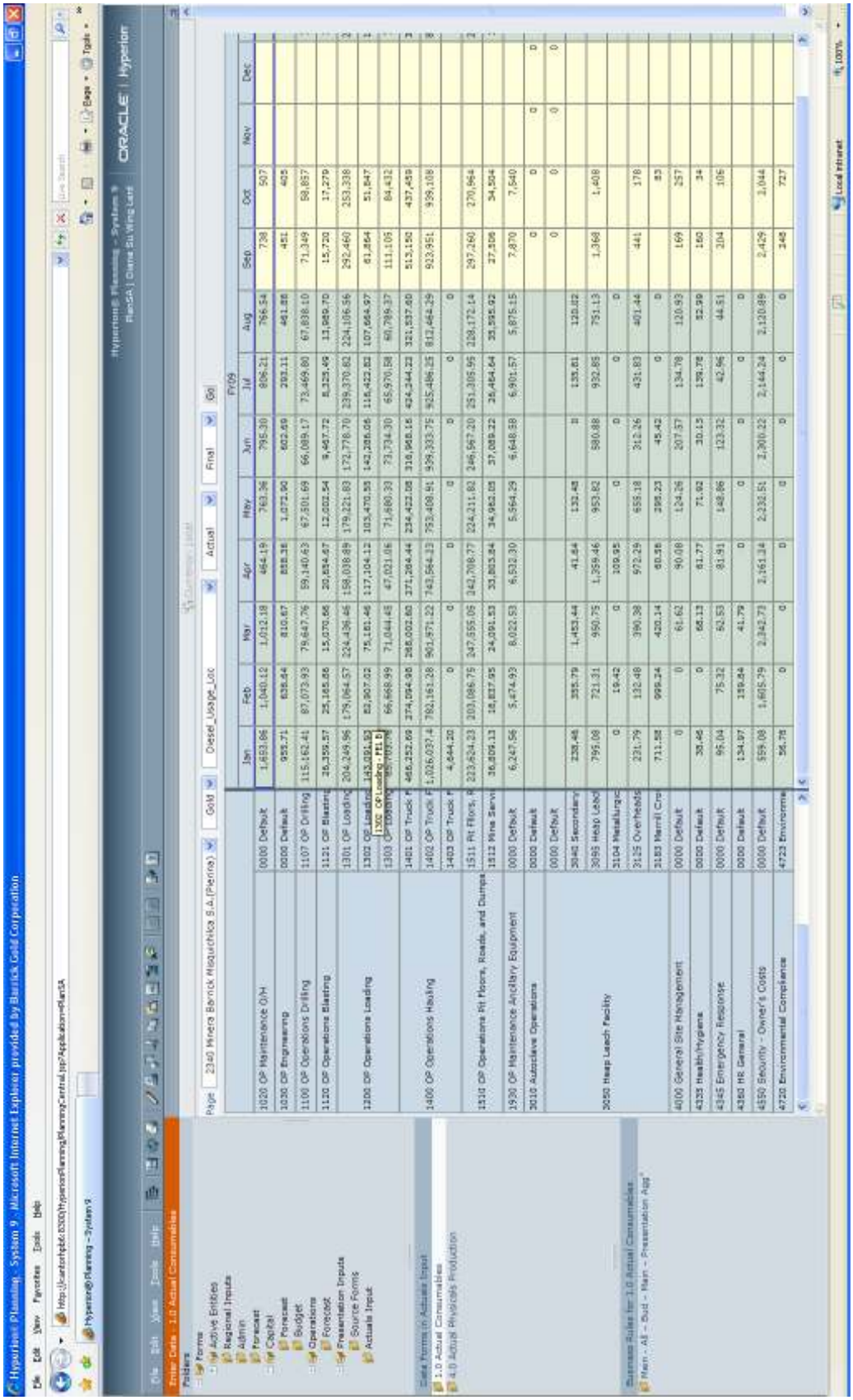


Figura 3.4 Sistema Hyperion Planning - System 9.0. Fuente: Minera Barrick Misquichilca. 2009.

3.7.3 J.D. Edwards de Oracle

J.D. Edwards es una compañía de software fundada en marzo de 1977 en Denver por Jack Thompson, Dan Gregory y Ed McVaney. Tuvo éxito creando un programa de contabilidad para los miniordenadores de IBM. La compañía fue añadiendo funciones, su software de contabilidad se convirtió en una aplicación E.R.P.independiente de la plataforma que en 1996 se llamó OneWorld.

En junio de 2003, el consejo de administración de J.D. Edwards accedió a la oferta de adquisición de PeopleSoft, completándose la adquisición en julio. A finales de 2004, PeopleSoft fue adquirida a su vez por Oracle.

Oracle JD Edwards World, creado para plataformas IBM Serie, ofrece a las pequeñas empresas un entorno confiable, rico en funciones y en ambiente web para la administración de alta calidad de plantas, inventarios, equipos, finanzas y personas como un todo integrado y sincronizado, que están previamente incorporados en una sola base de datos, reduciendo los costos y la complejidad de implementación.

JD Edwards World pertenece a la línea de productos de Aplicaciones Oracle, que también incluye PeopleSoft Enterprise, JD Edwards World EnterpriseOne e-Busines Suite.

JD Edwards tiene las siguientes familias de productos [Edwards01]:

- *Asset Lifecycle Management:* Control directo del desempeño de la empresa para obtener su máximo provecho.

- *Customer Relationship Management:* Soporta y optimiza el proceso del negocio desde el contacto original con el cliente hasta el servicio postventa.

- *Financial Management:* las soluciones de JD Edwards EnterpriseOne Financial Management están previamente integradas y se conectan sin dificultades con todas lás demás soluciones de JD Edwards EnterpriseOne.

- *Human Capital Management:* Ofrece un servicio más efectivo al cliente a través de las aplicaciones basadas en la Web, de autoservicio para los empleados y administradores.

- *Project Management:* Ofrece aplicaciones valiosas a los profesionales de proyectos vía Web.

- *Supply Chain Management:* Soporta procesos que promueven el crecimiento de ingresos, la reducción del inventario, una mejor utilización de activos y mejoras en el costo de productos al utilizar los mejores procesos de negocio

- *Supply Management (Procurement):* Administra su proceso de negocios a través de la colaboración en tiempo real durante el diseño, demanda, producción y planificación de distribución.

Los principales beneficios de JD Edwards World son:

- *Flexible y accesible:* Las aplicaciones previamente integradas y optimizadas sobre IBM Serie implican costos más bajos de implementación y necesidades actuales de IT. Posee una arquitectura flexible que permite adaptar los menús, la seguridad y el reporte respecto de las necesidades especificas de la empresa que lo adquiera sin modificaciones costosas.

- *Capacidades de autoservicio*: El acceso basado en Web browser a las aplicaciones permite a sus empleados, clientes y proveedores acceder a la información que les resulte relevante, de manera rápida y fácil con menos entrenamiento.

- *Solución completa y sólida:* JD Edward World es una solución integral y de bajo mantenimiento para pequeñas empresas. Ofrece la misma funcionalidad disponible para empresas más grandes, aunque no es una versión con las características básicas de una solución más grande. Soporta los requerimientos de múltiples divisas, múltiples empresas y múltiples idiomas y ofrece integración con otras tecnologías People Soft clave.

3.7.4 SAP ERP

SAP fue fundada en 1972 en la Ciudad de Mannheim, Alemania, por algunos de los antiguos empleados de IBM (Claus Wellenreuther, Hans-Werner Hector, Klaus Tschira, Dietmar Hopp y Hasso Plattner) bajo el nombre de SAP Sistemas, Applicaciones y Productos en Aplicaciones de Datos (Systemanalyse, Anwendungen und Programmentwicklung, original en alemán). El nombre fue tomado de la división en la que trabajaban en IBM. Su visión fue desarrollar un software estándar para el procesamiento del negocio en tiempo real [Sap01].

La corporación se desarrolla hasta convertirse en la quinta más grande compañía mundial de software. El nombre SAP R/3 es al mismo tiempo el nombre de una empresa y el de un sistema informático. Este sistema comprende muchos módulos completamente integrados, que abarca prácticamente todos los aspectos de la administración empresarial. Ha sido desarrollado para cumplir con las necesidades crecientes de las organizaciones mundiales. SAP ha puesto su mirada en el negocio como un todo: así ofrece un sistema único que soporta prácticamente todas las áreas en una escala global. SAP proporciona la oportunidad de sustituir un gran número de sistemas independientes, que se han desarrollado e instalado en organizaciones ya establecidas, por un solo sistema modular. Cada módulo realiza una función diferente, pero está diseñado para trabajar con otros módulos. Está totalmente integrado, ofreciendo real compatibilidad a lo largo de las funciones de una empresa.

Después de haber dominado el mercado, la empresa afronta una mayor competencia de Microsoft e IBM. En marzo de 2004 cambió su enfoque de negocio en favor de crear la plataforma que desarrolla y utiliza, la nueva versión de su software NetWeaver.

Es en este punto donde SAP se encuentra enfrentada con Microsoft e IBM, en lo que se conoce como "la guerra de las plataformas". Microsoft ha desarrollado una plataforma basada en la Web llamada .NET, mientras que IBM ha desarrollado otra llamada WebSphere.

A comienzos de 2006 fue anunciada una alianza muy importante entre SAP y Microsoft para integrar las aplicaciones ERP de SAP con las de Office de Microsoft bajo el nombre de proyecto "Duet".

La compra de SAP por parte de Microsoft habría sido uno de los acuerdos más grandes en la historia de la industria del software, dado el valor de mercado de la alemana, de más de 55.000 millones de euros (junio de 2004).

SAP ha conquistado clientes de forma consistente para aumentar la cuota del mercado global entre sus cuatro principales competidores a un 55% a fines de 2004, desde un 48% dos años antes. La participación combinada de Oracle y PeopleSoft declinó de un 29% a un 23%.

SAP está compuesto por módulos de aplicación, los cuales son programas individuales que pueden ser adquiridos, instalados y usados separadamente, pero todos extraen data de una misma base de datos, entre ellos se tienen:

- *Gestión Financiera (FI):* Libro mayor, libros auxiliares, ledgers especiales.
- *Controlling (CO):* Gastos generales, costes de producto, cuenta de resultados, centros de beneficio.
- *Tesorería (TR):* Control de fondos, gestión presupuestaria, flujo de efectivo.
- *Sistema de proyectos (PS):* Grafos, contabilidad de costes de proyecto.
- *Gestión de personal (HR):* Gestión de personal, cálculo de la nómina, contratación de personal.
- *Business Warehouse (BW) ó Business Intelligence (BI):* Data warehousing.

Las principales ventajas de SAP son:

- Fácil integración global. Se eliminan las barreras de tipo de cambio monetario, lenguaje y culturales.
- Las actualizaciones son necesarias sólo al momento de la implementación del ERP en la organización.
- Información en tiempo real, reduciendo la posibilidad de errores por redundancia.
- Agradable ambiente de trabajo haciendo que los empleados realicen su trabajo con facilidad incrementando su eficiencia.

- No se requiere comprar hardware y tampoco tiene costo de mantenimiento.

- No hay costos de entrenamiento para desarrolladores, sino que se capacita directamente a los usuarios finales.

- Provee acceso inmediato a la información de la organización.

Las principales desventajas de SAP son:

- El cerrar contrato con un vendedor de SAP antes de que éste expire para cambiar de proveedor puede no ser rentable.

- La personalización del SAP en una organización puede ser muy cara al no encajar con el modelo del negocio que lo implementa.

- El retorno de la inversión puede tomar mucho tiempo hasta que sea provechoso.

- La implementación del SAP ERP tiene un alto riesgo de que el proyecto fracase.

CAPÍTULO 4: Aporte Teórico

La elección del mejor algoritmo para una tarea empresarial específica puede ser un gran desafío. Aunque se pueden utilizar diferentes algoritmos para realizar la misma tarea, cada uno de ellos genera un resultado diferente. No es necesario usar los algoritmos de modo independiente, en una solución de minería de datos se puede usar un algoritmo para explorar los datos y luego usar otro algoritmo para predecir un resultado específico de esos datos.

Los modelos de minería de datos pueden predecir valores, generar resúmenes de datos y buscar correlaciones ocultas. En [Microsoft01] se muestra la Tabla 4.1 que sugiere los posibles algoritmos a usar para la tarea específica a desarrollar en minería de datos.

Tarea	Algoritmos que se pueden usar
Predecir un atributo discreto. Por ejemplo, predecir si el destinatario de una campaña de envío de correo directo adquirirá un producto.	Algoritmo de árboles de decisión Algoritmo Bayes Naive Algoritmo de clústeres Algoritmo de red neuronal
Predecir un atributo continuo. Por ejemplo, prever las ventas del año próximo.	Algoritmo de árboles de decisión Algoritmo de serie temporal
Predecir una secuencia. Por ejemplo, realizar un análisis clickstream del sitio web de una empresa.	Algoritmo de agrupación en clústeres de secuencia
Buscar grupos de elementos comunes en las transacciones. Por ejemplo, utilizar el análisis de la cesta de la compra para sugerir a un cliente la compra de productos adicionales.	Algoritmo de asociación Algoritmo de árboles de decisión
Buscar grupos de elementos similares. Por ejemplo, segmentar datos demográficos en grupos para comprender mejor las relaciones entre atributos.	Algoritmo de clústeres Algoritmo de agrupación en clústeres de secuencia

Tabla 4.1 Algoritmos de Minería de Datos que se pueden usar según la tarea a realizar. Fuente: [Microsoft01]

Debido a que cada modelo devuelve un tipo de resultado diferente, Analysis Services proporciona un visor independiente para cada algoritmo. Cuando se examina un modelo de minería de datos en Analysis Services, el modelo se muestra en la ficha *Visor de modelos de minería de datos*, que usa el visor adecuado para el modelo.

Como el objetivo de la presente aplicación es el de predecir un atributo continuo (indicadores o ratios de consumos), en la Tabla 4.1 podemos ubicar a nuestra solución seleccionando a los algoritmos de árboles de decisión o de series temporales.

4.1 Selección de la Solución para la Automatización del Proceso de Presupuestación

Para poder implementar la aplicación se han evaluado las diferentes técnicas que podían resolver el problema, la selección de la mejor técnica para automatizar el proceso de presupuestación está sustentada en comparaciones de distintos atributos o indicadores.

4.1.1 Atributos

En [Michalski98] se definen los siguientes atributos que miden el rendimiento o performance de las técnicas de minería de datos:

- **Precisión.** Mide la capacidad del algoritmo de llegar al resultado correcto. En definitiva, tratamos de evaluar el grado de error cometido en la respuesta. Algunas veces es importante distinguir entre dos tipos de errores: los ejemplos positivos clasificados como negativos (errores de omisión) y viceversa (errores de comisión); estos dos tipos de errores nos ayudan a determinar si los conceptos aprendidos son demasiado generales o demasiado específicos.

- **Claridad.** Mide la transparencia de una técnica, en cuanto a la interpretación de los posibles resultados, así como la forma en que se ha llegado a los mismos. Una red neuronal es un claro ejemplo de técnica que genera modelos de caja negra, en los cuales es difícil conocer cómo se producen las transformaciones internas que hacen llegar al resultado final.

- **Utilidad o Comprensibilidad.** La información tiene un valor que decrece con el paso del tiempo, por ello, resulta necesario que la técnica empleada genere información en un formato fácil de entender, con el objetivo de que tras la interpretación de quien tome las decisiones se convierta en un recurso de acción para la empresa. En definitiva, este indicador mide la forma de presentación de los resultados. Esto se mejora con el empleo de técnicas de visualización, de jerarquización de reglas, etc. Es importante que los conceptos generados sean comprensibles al usuario, ya que el fin último de estos sistemas es que el usuario aprenda algo de ellos.

- **Generalidad.** Se refiere a la posibilidad de aplicar la técnica a múltiples tipos de problemas, afectados por un amplio grupo de variables y con el empleo de varios tipos de datos.

- **Facilidad de Construcción.** La construcción del modelo suele ser bastante autónoma con relación a quien tome las decisiones. Sin embargo, también se ha de medir el coste de consecución en base a otros elementos, como el consumo de registros que necesita para el entrenamiento.

- **Gestión de Memoria.** Mide la necesidad de recursos de computación que son necesarios para que la herramienta pueda actuar correctamente. Generalmente, rapidez y recursos necesarios van de la mano, es decir, aquellos algoritmos más lentos son los que necesitan más capacidad de computación. Estos dos últimos criterios están reduciendo su importancia gracias a los avances que cada día se producen en las tecnologías de la computación.

- **Robustez.** ¿En qué medida es capaz el algoritmo de trabajar con datos perdidos o con errores, sin que afecte significativamente alcanzar el resultado óptimo? Se trata de un atributo relacionado con el preproceso, puesto que reduce la necesidad de éste. Es un atributo contra el ruido y contra los ejemplos incompletos, cada sistema maneja estos dos problemas de forma diferente, con lo cual debe evaluarse en cada sistema en particular.

- **Validación.** Se refiere a la facilidad para comprobar que el modelo ha llegado a la solución óptima. Así, las técnicas estadísticas disponen de buenos indicadores, como el coeficiente de determinación (R2). Habitualmente cada

herramienta construye sus propios indicadores, o se emplea la validación cruzada.

- **Disponibilidad.** Algunas técnicas están más disponibles en los distintos paquetes comerciales que otras. Así, las redes neuronales y los árboles de decisión son algoritmos usuales, mientras que los algoritmos genéticos difícilmente se encuentran.

4.1.2 Cuadro Comparativo de Soluciones

Con la finalidad de mostrar las ventajas y/o desventajas de las técnicas predictivas de regresión existentes en Minería de Datos, se ha elaborado el cuadro que compara a las series temporales y a los árboles de decisión, para ello se han considerado los atributos definidos anteriormente tomando en cuenta los objetivos del presente trabajo de investigación.

Atributos	Series Temporales	Árboles de Decisión
Precisión	No es necesario un gran número de datos para lograr una alta precisión en los resultados.	Las reglas de asignación son bastante sensibles a pequeñas perturbaciones en los datos (inestabilidad). Dificultad para elegir el árbol óptimo.
Claridad	No es intuitivo llegar al resultado en una serie temporal.	La forma de un árbol es intuitiva, el usuario puede comprobar la racionalidad del modelo.
Utilidad o Comprensibilidad	Presentan resultados fáciles de entender y usar por el usuario.	Son una excelente herramienta para manejar y disponer la información, pero presenta dificultades para clasificar las posibles combinaciones de datos en la predicción de series de tiempo. Es probable que el porcentaje de datos escogido para la construcción del árbol no sea el más competente para clasificar sus posibles combinaciones.
Generalidad	Predicen variables continuas.	Predicen variables discretas y continuas.
Facilidad de Construcción	Puede predecir tendencias basadas en el conjunto de datos original utilizado para crear el modelo.	Requieren columnas adicionales de nueva información como entrada para predecir una tendencia.
Robustez	Es robusto frente a datos atípicos.	No está indicado para trabajar con información incompleta
Validación	Es fácil comprobar que se ha llegado a la solución deseada.	Es válida sea cual fuera la naturaleza de las variables explicativas: continuas o discretas. Sin embargo, requieren de un gran número de datos para asegurarse que la cantidad de las observaciones de los nodos hoja es significativa.

Tabla 4.2 Cuadro comparativo de soluciones. Fuente: Elaboración propia

De acuerdo a las características mencionadas en la Tabla 4.2, se les puede calificar de acuerdo a los valores mostrados en la Tabla 4.3:

Puntaje	Valor
1	Bueno
0.5	Regular
0	Malo

Tabla 4.3 Tabla de valores de calificación. Fuente: Elaboración propia

Una vez definido este puntaje, procederemos a realizar las comparaciones entre cada algoritmo, como se indica en la Tabla 4.4:

Atributos/Algoritmos	Series Temporales	Árboles de Decisión
Precisión	1	0.5
Claridad	0	1
Utilidad o Comprensibilidad	1	0.5
Generalidad	0.5	1
Facilidad de Construcción	1	0.5
Robustez	1	0
Validación	1	0.5
Puntaje Final	**5.5**	**4**

Tabla 4.4 Tabla de Calificación de los algoritmos predictivos regresivos. Fuente: Elaboración propia

Como podemos observar en la Tabla 4.4, el algoritmo de series temporales es la solución seleccionada para la optimización del proceso de presupuestación, es la técnica que está más alineada a los objetivos que se buscan implementar en el presente proyecto de investigación.

4.2 Comparación de Metodologías de desarrollo de proyectos de Minería de Datos

Las metodologías SEMMA y CRISP-DM comparten la misma esencia, estructurando el proyecto de Minería de Datos en fases que se encuentran interrelacionadas entre sí, convirtiendo el proceso de Minería de Datos en un proceso iterativo e interactivo. La metodología SEMMA se centra más en las características técnicas del desarrollo del proceso, mientras que la metodología CRISP-DM, mantiene una perspectiva más amplia respecto a los objetivos empresariales del proyecto. Esta diferencia se establece ya desde la primera fase del proyecto de Minería de Datos donde la metodología SEMMA comienza realizando un muestreo de datos, mientras que la metodología CRISP-DM comienza realizando un análisis del problema empresarial para su transformación en un problema técnico. Desde ese punto de vista más global se puede considerar que la metodología CRISP-DM está más cercana al

concepto real de proyecto, pudiendo ser integrada con una Metodología de Gestión de Proyectos específica que completaría las tareas administrativas y técnicas.

Otra diferencia significativa entre la metodología SEMMA y la metodología CRISP-DM radica en su relación con herramientas comerciales. La metodología SEMMA sólo es abierta en sus aspectos generales ya que está muy ligada a los productos SAS donde se encuentra implementada. Por su parte la metodología CRISP-DM ha sido diseñada como una metodología neutra respecto a la herramienta que se utilice para el desarrollo del proyecto de Minería de Datos siendo su distribución libre y gratuita.

Característica	SEMMA	CRISP-DM
Permite elección totalmente libre de herramientas	No	Sí
Cantidad de fases	5	6
Todas las fases pueden relacionarse	Sí	Sí
Considera otros aspectos no técnicos	No	Sí
Está detallado paso a paso cada etapa	No	Sí

Tabla 4.5 Comparación entre metodologías de proyectos de Minería de Datos. Fuente: Elaboración propia

Según el resultado obtenido en la Tabla 4.5, se puede concluir que la metodología CRISP-DM es la que ofrece más ventajas en comparación con la metodología SEMMA, puesto que permite elegir libremente las herramientas a usarse en la aplicación de Minería de Datos y tiene todas las etapas detalladas paso a paso.

4.3 Comparación de Herramientas de Inteligencia de Negocios

Para realizar la comparación de las principales herramientas de Inteligencia de Negocios nos vamos a basar en el Cuadrante Mágico para Plataformas de Inteligencia de Negocios (Magic Quadrant for Business Intelligence Platforms) de Gartner [Gartner01], el cual presenta una visión global de la opinión de Gartner acerca de los principales fabricantes de software que deben ser considerados por las organizaciones que buscan desarrollar la inteligencia de negocios (BI) dentro de sus aplicaciones. Los compradores deben evaluar a los proveedores en los cuatro cuadrantes y no asumir que sólo las organizaciones más grandes pueden ofrecer éxito en las implementaciones de BI. Además de las opiniones de los analistas de Gartner, las puntuaciones y comentarios en este documento se basan en tres fuentes: la percepción del cliente acerca de cada uno de los puntos fuertes de cada vendedor y los retos derivados de la BI-relacionadas con las investigaciones de Gartner, una encuesta en línea de proveedores a los clientes llevada a cabo a finales de 2008, de

las cuales se obtuvieron 480 respuestas, y un cuestionario completo al proveedor acerca de su estrategia de BI y sus operaciones.

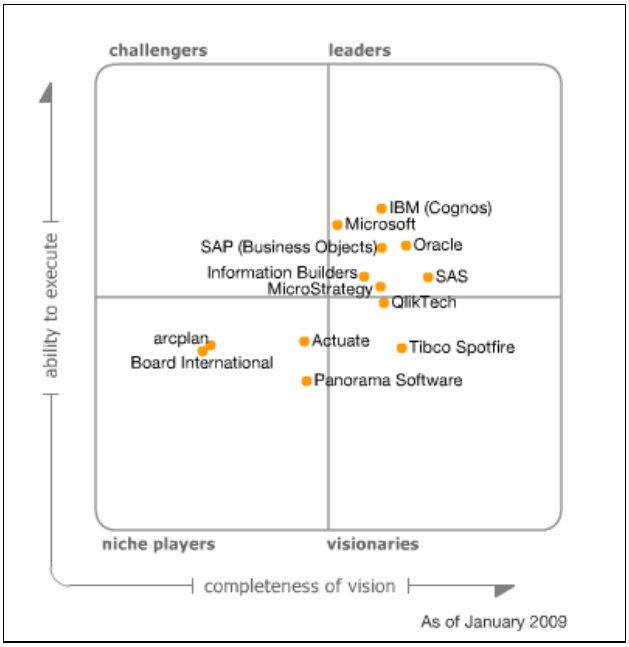


Figura 4.1 Cuadrante Mágico de Gartner para las plataformas de BI. Fuente: [Gartner01]

4.3.1 Criterios de evaluación

1. Habilidad para ejecutar

Los vendedores son calificados por su habilidad y éxito en hacer que su visión de mercado se haga realidad.

- Producto/Servicio: ¿Qué tan competitivo y de éxito son los bienes y servicios ofrecidos por el proveedor en este mercado?

- Viabilidad: ¿Cuál es la probabilidad de que el proveedor siga invirtiendo en productos y servicios para sus clientes?

- Ejecución de Ventas/Precios: ¿Tiene el proveedor la capacidad de proporcionar costos eficaces de licencias y opciones de mantenimiento?

- Respuesta de mercado y Track Record: ¿Puede el vendedor responder a los cambios en la dirección del mercado como los requisitos de los clientes evolucionan?

- Mercado de ejecución: ¿Están los clientes al tanto de la oferta del proveedor en el mercado

- La experiencia del cliente: ¿Qué tan bien el proveedor da asistencia técnica a sus clientes?

- Operaciones: ¿Cuál es la capacidad de la organización para alcanzar sus metas?

2. Integridad de la Visión

Los vendedores se han centrado en comprender cómo las fuerzas del mercado pueden ser aprovechadas para crear valor para sus clientes y oportunidades para ellos.

- Comprensión del Mercado: ¿El proveedor tiene la capacidad de comprender las necesidades de los compradores, y de traducir esas necesidades en productos y servicios?

- Estrategia de Marketing: ¿Tiene el proveedor un conjunto claro de los mensajes que comunican su valor y diferenciación en el mercado?

- Estrategia de Ventas: ¿El proveedor tiene la combinación correcta de los recursos directos e indirectos para ampliar su alcance en el mercado?

- Ofreciendo (producto) Estrategia: ¿El enfoque del proveedor para el desarrollo de productos y la prestación de destacar la diferenciación y la funcionalidad, se alinean a las necesidades actuales y futuras?

- Modelo de Negocio: ¿Qué tan firme y lógica es la proposición de negocios subyacentes del proveedor?

- Estrategia Vertical/Industria: ¿Qué tan bien el vendedor puede satisfacer las necesidades de diversas industrias, tales como servicios financieros o retail?

- Estrategia Geográfica: ¿Qué tan bien el vendedor puede satisfacer las necesidades de lugares fuera de su país de origen, ya sea directamente o a través de socios?

De acuerdo a los criterios de evaluación podemos deducir que ***Microsoft*** ofrece una mayor ventaja sobre otros proveedores disponibles en cuanto a Habilidad para Ejecutar. Y este criterio es el que nos interesa para nuestro estudio porque está más relacionado con el producto, el cual usaremos para implementar la presente aplicación.

4.3.2 Microsoft SQL Server 2008

SQL Server 2008 de Microsoft es una base de datos bastante completa, la cual se adapta a las necesidades de cualquier empresa independientemente de cual sea su tamaño. Gracias a su gestión integral de los archivos de una compañía, hace posible controlar cualquier información desde cualquier lugar y en cualquier momento.

La estructura de SQL Server 2008 gira en torno a tres plataformas distintas, las cuales se complementan para generar una solución que integra fiabilidad, productividad e inteligencia. La primera es la base de datos fiable. Su infraestructura permite el cifrado completo de la base de datos y la gestión de claves sin que sea necesario modificar en modo alguno las aplicaciones

La segunda plataforma es la productiva. Su función es mejorar el proceso de instalación y la capacidad de almacenamiento de datos. SQL Server 2008 incorpora mejoras en el proceso de vida de la base de datos, así como también un rediseño de los procesos de instalación y la arquitectura de configuración. Además cuenta con una gestión basada en políticas, lo que permite prevenir cambios no deseados en la configuración de uno o más servidores SQL Server y asegura el cumplimiento de las políticas impuestas por el administrador de bases de datos o la compañía.

La tercera pata sobre la que se sustenta el programa de Microsoft es la ***plataforma Business Inteligence***. Las herramientas que incluye se dirigen a la gestión de aquellos archivos que faciliten la toma de decisiones mediante la comprensión del funcionamiento de la firma y la anticipación de acciones para otorgarle una dirección bien informada. Para ello, integra sistemas de partición que posibilitan manejar tablas de gran tamaño de un modo más eficaz gracias a que las dividen en bloques más pequeños.

4.4 Diseño de la solución

Tomando en cuenta las metodologías y herramientas expuestas y elegidas en el presente capítulo, proponemos como solución, al problema expuesto, la implementación de una aplicación basada en Minería de datos haciendo uso de algoritmos de serie temporal para la optimización del proceso de presupuestación. Luego de un previo análisis, se ha visto por conveniente utilizar la herramienta de minería de datos del Analysis Services de Microsoft SQL Server 2008. Adicionalmente, se utilizará Visual Studio 2005 para el desarrollo de la aplicación por tener una adecuada compatibilidad con Microsoft SQL Server 2008.

El procesamiento de la información, haciendo uso de algoritmos de serie temporal, se realizará con el Analysis Services de MS SQL Server 2008, por ello nos centraremos básicamente en 2 procesos: el primero abarcará la **comprensión** y **preparación** de los datos históricos (que servirán de entrada para nuestro análisis de minería de datos), para lo cual utilizaremos la metodología de implementación de proyectos de minería de datos CRISP-DM, y el segundo concerniente a la **implementación** de la interfaz cliente, la que servirá para la carga de datos adicional, parametrizar los resultados a mostrar y generar un reporte personalizado con el fin de mejorar el proceso de obtención de datos que ayudarán a optimizar el proceso de presupuestación.

En la Figura 4.2 se muestra la arquitectura de la solución a implementar:

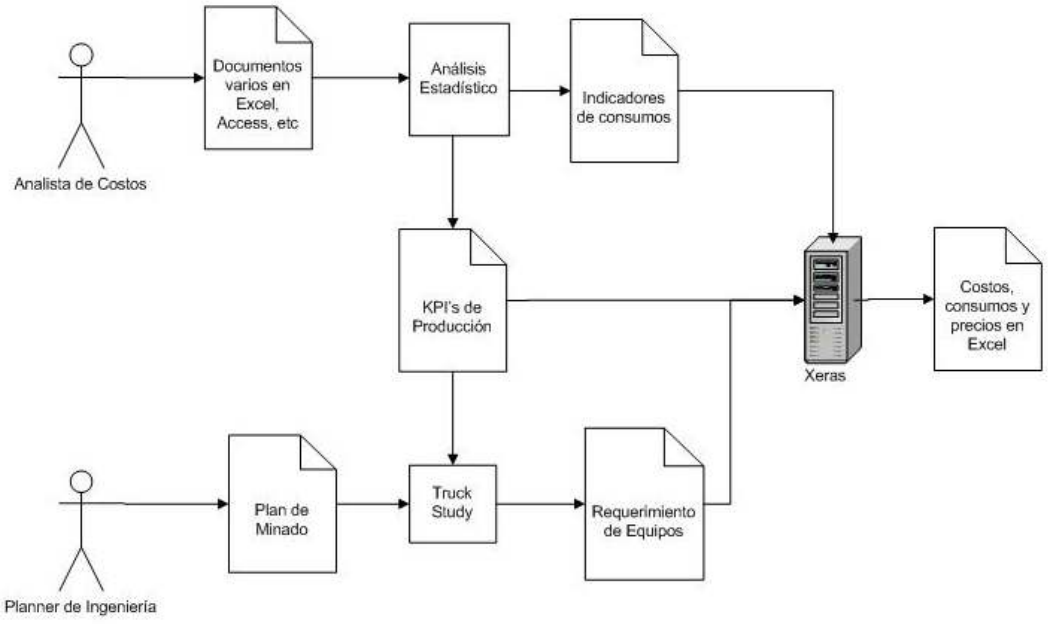


Figura 4.2 Arquitectura de la solución a implementar. Fuente: Elaboración propia

Para la construcción del sistema de proyección de costos basado en minería de datos, seguiremos los pasos descritos por la metodología de desarrollo de proyectos de minería de datos CRISP-DM, cuyas etapas son las siguientes:

1. *Comprensión del negocio:* Es la fase inicial, aquí nos enfocaremos en comprender los objetivos del proyecto y sus exigencias desde una perspectiva del negocio. En esta fase definimos el problema.

2. *Comprensión de los datos:* Se recolectarán los datos iniciales, habrá una familiarización con los datos, identificaremos los problemas de calidad de datos, descubrimiento de relaciones iniciales que formarán hipótesis en cuanto a la información oculta.

3. *Preparación de los datos:* El objetivo es construir el conjunto de datos final, a partir de los datos puros iniciales. Las tareas a desarrollar serán: selección de tablas, registros y atributos, así como también la transformación y limpieza de datos para las herramientas que se encargarán de modelar. Adicionalmente, en esta etapa, nos encargaremos de poblar y procesar nuestro Datamart.

4. *Modelamiento:* En esta etapa, elegiremos y aplicaremos más de una técnica de modelado. Por lo general, existen varias técnicas para un mismo problema de minería de datos. Algunas de estas técnicas tienen requerimientos específicos sobre la forma de los datos.

5. *Evaluación:* Al llegar a esta etapa, ya se cuenta con un modelo adecuado para realizar un análisis de datos. Antes de proceder al despliegue final del modelo, es importante revisar a fondo los pasos que se han seguido para crearlo, para comparar el modelo correctamente obtenido con los objetivos de negocio.

6. *Explotación:* Es la obtención del conocimiento a partir del modelo obtenido, este conocimiento ganado tendrá que ser organizado y presentado de una manera sencilla de utilizar por el usuario final. Dependiendo de los requerimientos, la fase de desarrollo puede ser tan simple como la generación de un informe o tan compleja como la realización repetida de un proceso cruzado de minería de datos a través de la organización. Esta fase incluye la implementación del aplicativo, es decir, el sistema de proyección de costos.

7. *Validación de la solución:* Se realiza una simulación completa del sistema para confirmar que se ha cumplido con el objetivo propuesto, la cual se presentará en el capítulo siete.

CAPÍTULO 5: Aporte Práctico

En este capítulo se va a describir la metodología que pretende descubrir patrones repetitivos y sus interdependencias dentro de series temporales de datos históricos en el proceso de presupuestación.

5.1 Caso de Estudio

Minera Barrick Misquichilca S.A. es una empresa perteneciente al sector de la gran minería dedicada a la extracción de oro y plata. En Perú cuenta con 2 unidades mineras ubicadas en Huaraz (Pierina) y en la sierra de La Libertad (Lagunas Norte), además de su sede administrativa ubicada en Lima. Nuestro estudio se centra en el área de Operaciones Mina de la sede minera Pierina. Para mayor detalle del presente caso estudio ver Anexo.

5.2 Propuesta de Solución

En el presente proyecto de investigación vamos a seguir los lineamientos de la metodología CRISP-DM, donde algunas fases ya se encuentran definidas en capítulos anteriores.

5.2.1 Comprensión del Negocio

Nos enfocamos a comprender los objetivos del proyecto y exigencias desde una perspectiva del negocio. En esta fase definimos el problema.

Análisis del Problema

Los analistas de costos del área de Operaciones Mina de Minera Barrick Misquichilca elaboran presupuestos de costos mensuales y/o anuales en base a ciertos indicadores, información que no se encuentra integrada, lo cual convella a que los analistas tengan que hacer cálculos manuales para obtener esta información. Los analistas tienen a su disposición un Datamart de Costos que contiene la información histórica de sus gastos y/o costos, sin embargo, esta información no es explotada.

En base a lo expuesto, se definió la problemática y se propone la siguiente solución:

Predecir los costos en base a la información residente en el Datamart de Costos, para lo cual se usará Minería de Datos y se implementará una aplicación, la cual permitirá a los usuarios explotar esta información, logrando alcanzar de esta manera nuestros objetivos planteados: obtener datos más exactos y confiables e integrar las diversas fuentes de datos, optimizando de esta manera el proceso de presupuestación.

5.2.2 Comprensión de los Datos

Luego de analizar el problema, la tarea principal de esta etapa es la descripción de los datos.

La fuente principal de los datos han sido los reportes generados por los analistas de costos en períodos pasados, estos corresponden a los años 2007, 2008 y 2009. También se obtiene información del Datamart de Costos con que cuentan en la actualidad los usuarios, a continuación presentamos la Tabla 5.1 que muestra el volumen de cada origen de datos:

Fuente	Porcentaje
Documentos de trabajo	80%
Actual Datamart de Costos	15%
Otros	5%

Tabla 5.1 Volumen de datos por fuente de origen de datos. Fuente: Elaboración propia

Toda esta información recolectada de las diversas fuentes de datos es almacenada en la base de datos *Presupuestos* mediante un proceso de carga de

datos. La base de datos ha sido implementada en el motor de base de datos de Microsoft SQL Server 2008. En la figura 5.1 se muestra el esquema de implementación de la integración de fuentes.

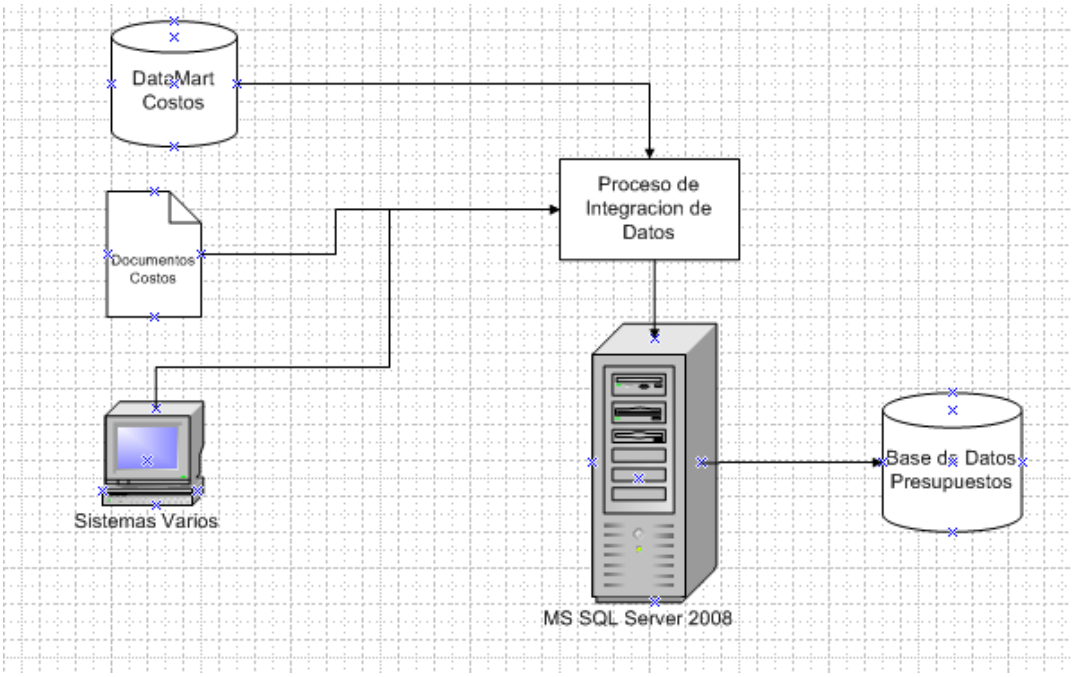


Figura 5.1 Esquema de integración de fuentes de datos. **Fuente: Elaboración propia**

Luego de extraer y almacenar la data en la base de datos Presupuestos, se han procedido a analizar y seleccionar los atributos más importantes y significativos para nuestra aplicación, de acuerdo a los siguientes criterios:

- **Dimensión de tiempo:** los analistas de costos realizan un presupuesto anual en base a períodos de tiempo mensuales.

- **Criterio de evaluación:** el principal criterio de evaluación es el ratio de consumo de los principales indicadores de costos en el área de Operaciones Mina, tales como: galones de diesel por hora para cada equipo minero, tonelada de emulsión matriz por tonelada de material volado, tonelada de nitrato de amonio por tonelada de material volado.

- **Variables de decisión:** teniendo en cuenta que el objetivo del proceso de presupuestación consiste en la proyección de costos, las variables de decisión consideradas son las siguientes:

- Consumo de diesel.
- Consumo de nitrato de amonio.
- Consumo de emulsión matriz.

Para nuestra aplicación práctica se ha tomado en cuenta la variable de consumo de diesel. Estas variables serán proyectadas a un cierto período de tiempo (mensual o anualmente).

- **Parámetros numéricos:** estos parámetros son las variables de entrada que sirven para encontrar los patrones de comportamiento en nuestra solución de minería de datos. Tenemos los siguientes parámetros:
  - Horas operativas de cada equipo minero.
  - Consumo de diesel
  - Tonelaje volado.
  - Explosivos (nitrato de amonio y emulsión matriz) usados por tonelada volada.

  Para nuestra aplicación práctica se ha tomado en cuenta como parámetros de entrada las horas operativas y el consumo de diesel.

Los atributos más importantes para el proyecto de Minería de Datos serán analizados en el diseñador de vistas de origen de datos de Business Intelligence Development Studio y el editor de consultas del Management Studio. Con estas herramientas se estudiarán los atributos, sus valores y el comportamiento de los mismos.

5.2.3 Preparación de los datos

En esta etapa hemos procedido a seleccionar un subconjunto de los datos adquiridos en la fase previa, basándonos en las características recalcadas en dicha fase, creamos conjuntos de datos válidos, para luego aplicar las técnicas de prospección en la fase siguiente de modelado.

Representación de los datos de entrada (Input Data)

En la mayoría de los casos, la entrada (Input Data) de un análisis de minería de datos toma la forma de una tabla de dos dimensiones, llamado “conjunto de datos” (dataset), independientemente de la representación lógica y física adoptadas para almacenar la información en archivos, bases de datos, datawarehouses y datamarts

utilizados como fuentes de datos. Los registros en el conjunto de datos seleccionado corresponden a las observaciones registradas en el pasado, los cuales también se conocen como ejemplos, casos, instancias o registros. En nuestro modelo los datos corresponden a información histórica de los 3 últimos años (2007-2009).

Las columnas representan la información disponible para cada observación y se denominan atributos, variables, características o funciones. Toda la información se encuentra organizada y disponible en el Datamart de Presupuestos. Siguiendo los criterios descritos en la anterior etapa, en la Tabla 5.2 se muestra el conjunto de datos seleccionado:

PERIODO	COMPAÑÍA	CENTRO COSTO	ACTIVIDAD	CUENTA	ITEM	DESCRIPCIÓN	UNIDAD	DESCRIPCIÓN	CANTIDAD TOTAL	HORAS OPERATIVAS	CANTIDAD HORARIA	COSTO UNITARIO
1	2340	1200	1301	56210	614	DIESEL	2	Galon	14473	230.4	62.82	2.79
2	2340	1200	1301	56210	614	DIESEL	2	Galon	34192	549.4	62.24	2.79
3	2340	1200	1301	56210	614	DIESEL	2	Galon	35441	532.3	66.58	2.79
4	2340	1200	1301	56210	614	DIESEL	2	Galon	27786	453.5	61.27	2.79
5	2340	1200	1301	56210	614	DIESEL	2	Galon	38102	579.3	65.77	2.79
6	2340	1200	1301	56210	614	DIESEL	2	Galon	40243	610.4	65.93	2.79
7	2340	1200	1301	56210	614	DIESEL	2	Galon	31937	499.9	63.89	2.79
8	2340	1200	1301	56210	614	DIESEL	2	Galon	27060	424	63.82	2.79
9	2340	1200	1301	56210	614	DIESEL	2	Galon	23841	390.1	61.12	2.79
10	2340	1200	1301	56210	614	DIESEL	2	Galon	36610	638.6	57.33	2.79
11	2340	1200	1301	56210	614	DIESEL	2	Galon	33014	552.9	59.71	2.79
12	2340	1200	1301	56210	614	DIESEL	2	Galon	31854	563.4	56.54	2.79
13	2340	1200	1301	56210	614	DIESEL	2	Galon	32222	559.8	57.56	3
14	2340	1200	1301	56210	614	DIESEL	2	Galon	8780	144.9	60.59	3.05
15	2340	1200	1301	56210	614	DIESEL	2	Galon	17938	315.9	56.78	3.17
16	2340	1200	1301	56210	614	DIESEL	2	Galon	27435	506.8	54.13	3.12
17	2340	1200	1301	56210	614	DIESEL	2	Galon	34802	599.1	58.09	3.08
18	2340	1200	1301	56210	614	DIESEL	2	Galon	27864	517.7	53.82	3.07
19	2340	1200	1301	56210	614	DIESEL	2	Galon	27804	437.5	63.55	3.24
20	2340	1200	1301	56210	614	DIESEL	2	Galon	30677	482.7	63.55	3.17
21	2340	1200	1301	56210	614	DIESEL	2	Galon	21245	334.3	63.55	3.21
22	2340	1200	1301	56210	614	DIESEL	2	Galon	32183	506.4	63.55	3.1
23	2340	1200	1301	56210	614	DIESEL	2	Galon	32037	504.1	63.55	3.1
24	2340	1200	1301	56210	614	DIESEL	2	Galon	5935	93.4	63.54	2.98
25	2340	1200	1301	56210	614	DIESEL	2	Galon	26512	445.2	59.55	2.73
26	2340	1200	1301	56210	614	DIESEL	2	Galon	17937	264.3	67.87	2.39
27	2340	1200	1301	56210	614	DIESEL	2	Galon	27323	456	59.92	2.47
28	2340	1200	1301	56210	614	DIESEL	2	Galon	21789	349.8	62.29	2.5
29	2340	1200	1301	56210	614	DIESEL	2	Galon	28283	428.3	66.04	2.57
30	2340	1200	1301	56210	614	DIESEL	2	Galon	11056	133.2	83	2.54
31	2340	1200	1301	56210	614	DIESEL	2	Galon	28238	264.3	106.84	2.55
32	2340	1200	1301	56210	614	DIESEL	2	Galon	34925	547.1	63.84	2.6
33	2340	1200	1301	56210	614	DIESEL	2	Galon	36397	558.4	65.18	2.64
34	2340	1200	1301	56210	614	DIESEL	2	Galon	37584	603.7	62.26	2.66
35	2340	1200	1301	56210	614	DIESEL	2	Galon	27950	484	57.75	2.65
36	2340	1200	1301	56210	614	DIESEL	2	Galon	27187	455.2	59.73	2.7

Tabla 5.2. Conjunto de datos seleccionado para la aplicación del modelo de minería de datos. **Fuente: Elaboración propia**

5.2.4 Modelamiento

En esta fase se seleccionamos las técnicas de modelado, las aplicamos sobre el conjunto de datos seleccionado y se calibran los parámetros a valores óptimos. Para la realización de esta fase se han utilizado las técnicas de Minería de Datos que vienen incluidas en el suite de soluciones de SQL Server Business Intelligence Development Studio de Microsoft SQL Server 2008, específicamente la herramienta SQL Server Analysis Services (SSAS).

Iniciamos el modelado creando un proyecto de Analysis Services usando la herramienta SQL Server Business Intelligence Development Studio. Una vez creado el proyecto, seguimos los siguientes pasos:

1. Creación del Origen de Datos
2. Creación Vista del Origen de Datos
3. Creación del Cubo del DataMart Presupuestos
4. Desarrollo de una Estructura de Minería de Datos

Este último paso nos permitirá analizar el modelo de acuerdo al algoritmo/técnica seleccionada, en la Tabla 5.3 se muestra la técnica utilizada de acuerdo a nuestro objetivo propuesto en el presente trabajo.

Objetivo de Minería	Técnica
Realizar una Proyección adecuada de los costos, tomando en cuenta costos de periodos pasados.	Algoritmo de Series Temporales de Microsoft

Tabla 5.3 Técnica a aplicar por objetivo de la minería de datos. Fuente: Elaboración propia

El detalle del desarrollo de esta fase se muestra en el Capítulo 6 del presente trabajo.

5.2.5 Evaluación

En esta fase se realizó el diseño de las pruebas sobre el conjunto de datos, el cual se realizó utilizando la herramienta SQL Server Integration Services, empleando la técnica de validación cruzada.

SQL Server Integration Services tiene componentes que permiten obtener muestras aleatorias representativas según un porcentaje de los datos o según

determinada cantidad de filas.

En esta fase se evaluó el modelo escogido desde el punto de vista del cumplimiento de los objetivos del negocio. Revisamos el proceso, teniendo en cuenta los resultados obtenidos, para repetir alguna fase en caso de que se hayan cometido errores. Evaluamos la validez del modelo generado, en función de los criterios de éxito establecidos en la primera fase y de la precisión del mismo.

Se validaron además los datos y parámetros numéricos ingresados en el modelo preliminarmente. Estos datos provienen del Datamart Presupuestos previamente implementado. Una vez que se obtuvieron los primeros resultados numéricos usando la solución propuesta, el modelo fue validado sometiendo las conclusiones a los analistas de costos que manejarán esta aplicación. En esta fase además se consideraron los siguientes factores:

- La probabilidad de alcanzar las conclusiones.

- La consistencia de los resultados en valores extremos de parámetros numéricos.

- La estabilidad de los resultados cuando se introducen cambios menores en los parámetros de entrada.

Resumen de evaluación de los resultados

A continuación se muestra la Tabla 5.4 con el porcentaje estimado de cumplimiento del objetivo del negocio basado en los criterios de éxito:

Criterios de éxito del negocio	Cumplimiento estimado
Obtener un modelo de conocimiento y comprobar que las conclusiones obtenidas son válidas o útiles	100%
Desarrollar el caso de estudio utilizando las herramientas de SQL Server 2008 para minería de datos	100%
Realizar un proyecto de Minería de Datos guiado por la metodología CRISP-DM y la documentación de cada una da las fases	100%
Interpretar los resultados de la relación que existe entre cada equipo y el ratio de consumo de los principales indicadores de costos	100%

Tabla 5.4 Estimado de cumplimiento de los criterios de éxito del negocio. Fuente: Elaboración propia

5.2.6 Explotación

Una vez creado el modelo de minería de datos, se ha visto por conveniente desarrollar una aplicación basada en el modelo de conocimiento obtenido con la finalidad de incrementar el conocimiento de los datos, este conocimiento obtenido necesita organizarse y representarse de manera que pueda ser usado. Con el sistema desarrollado mostramos un ejemplo de lo que se podría conseguir mediante la aplicación de procesos de prospección a datos de costos para optimizar el proceso de presupuestación, minimizando de esta manera los tiempos de obtención de información y mejorando la confiabilidad de los datos obtenidos.

En el capítulo 6 se describe la implementación del sistema de proyección de costos.

5.2.7 Validación de la Solución

En el capítulo 7 se realiza una recolección de datos, que serán analizados estadísticamente por el método de Montecarlo y se realizará una simulación de datos para validar la aplicación de series temporales desarrollada.

La variable elegida es el ratio *consumo horario de diesel* en un equipo minero cargador Frontal Komatsu WA1200-1, código 8A5002.

CAPÍTULO 6: Implementación de la Aplicación

En el presente capítulo se mostrará cómo se ha diseñado el Sistema de Proyección de Presupuestos, desde el diseño lógico de la Base de Datos "Presupuestos", el diseño del datamart "Presupuestos", el diseño del modelo de Minería de Datos (Orígenes de datos, Vistas del origen de datos, Estructuras de minería de datos y los modelos de minería de datos).

Asimismo, se detalla el modelado del negocio, que abarca la descripción de los actores y trabajadores del negocio, los casos de uso del negocio y el diagrama de actividades para cada caso de uso.

Finalmente, se muestra la interface gráfica de la aplicación, con la que el usuario tendrá la interacción para obtener los datos que requiere, según ciertos parámetros de entrada.

6.1 Diseño Lógico de la Base de Datos "Presupuestos"

En la Figura 6.1 se puede observar el diseño lógico de la base de datos "Presupuestos", donde se han creado tablas para almacenar los datos históricos relacionados con las cantidades, precios y gastos incurridos en consumibles por cada equipo minero, con periodicidad mensual, relacionados a los gastos del área de Operaciones Mina.

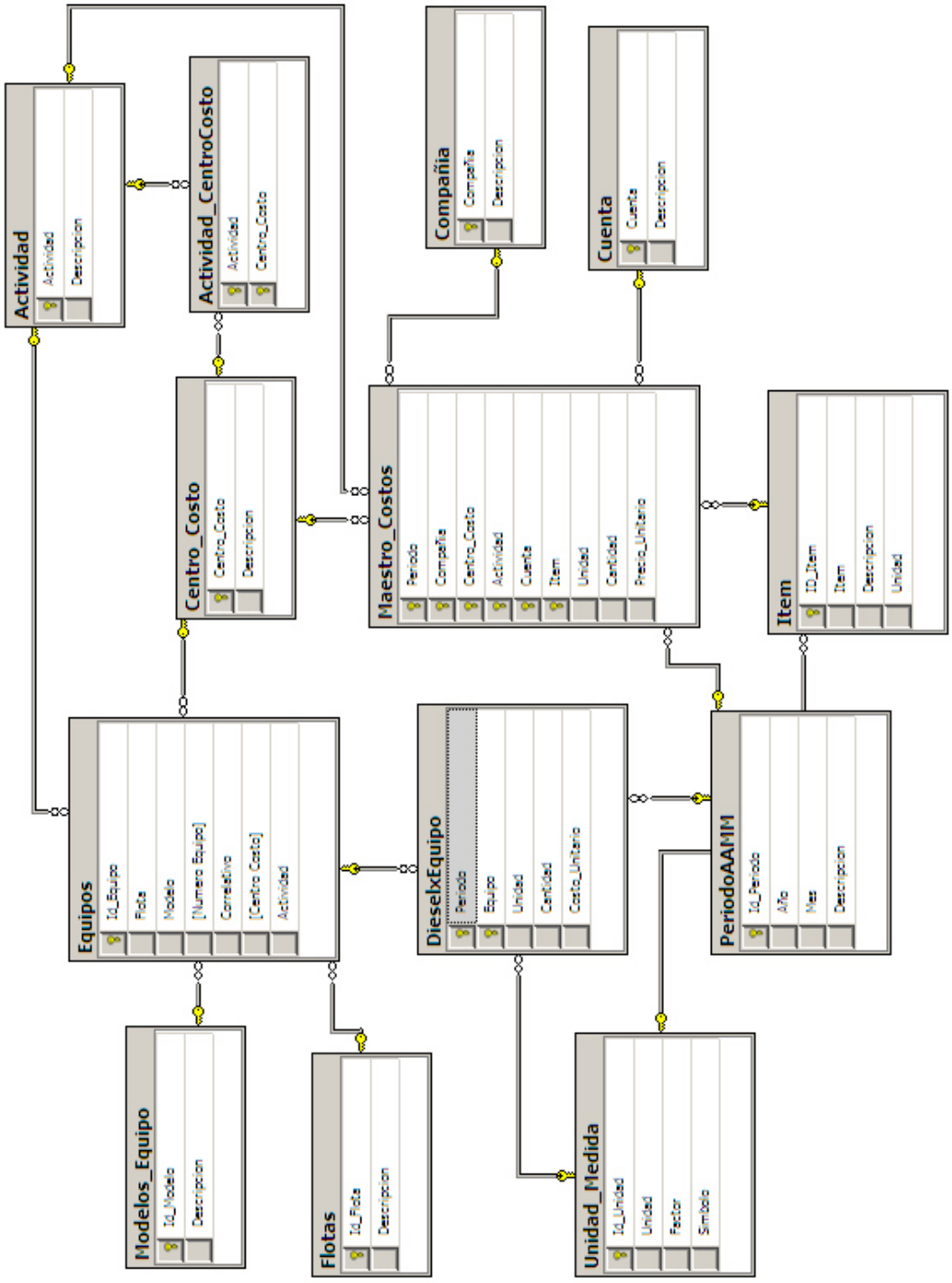


Figura 6.1 Diseño Lógico de la Base de Datos “Presupuestos”

6.2 Diseño del Datamart "Presupuestos"

Una vez que se ha cargado la información histórica en la Base de Datos "Presupuestos", se procede a diseñar el Datamart "Presupuestos" el cual estructura toda la información de modo que nos permita desarrollar más adelante la solución de Minería de Datos.

Iniciamos creando un proyecto de Analysis Services usando la herramienta de SQL Server Business Intelligence Development Studio del Microsoft SQL Server 2008.

A continuación, detallamos cada uno de los pasos realizados en el desarrollo del Datamart "Presupuestos" de nuestra aplicación:

6.2.1 Origen de Datos

Se establece la fuente de datos desde donde se va a obtener la información, en este caso es la base de datos Presupuestos. El origen de datos define la cadena de conexión e información de autenticación que el servidor Analysis Services utilizará para conectarse al origen de datos. El origen de datos puede contener varias tablas o vistas.

En la Figura 6.2 se muestra la configuración del origen de datos.

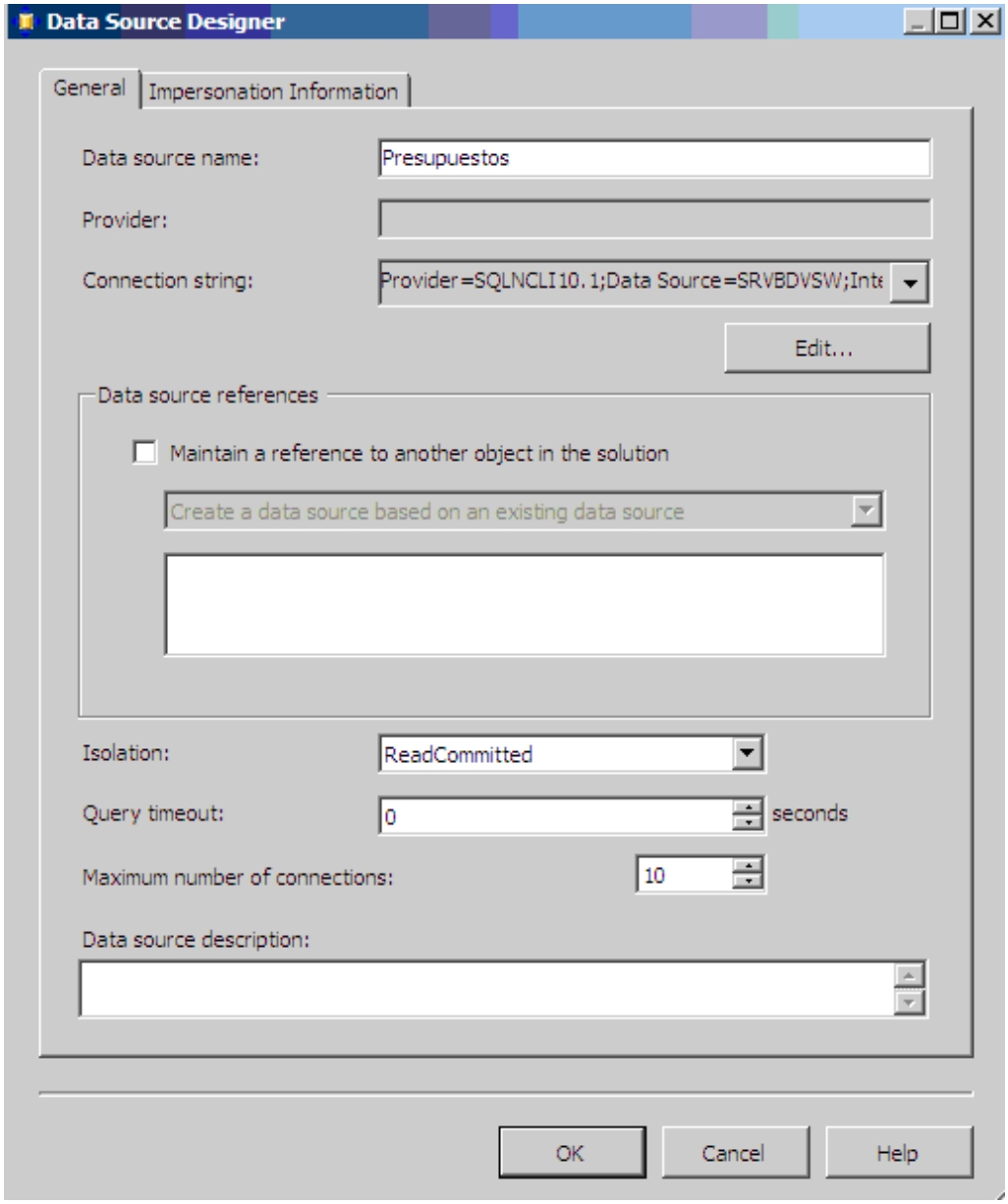


Figura 6.2 Configuración del Origen de Datos

6.2.2 Vista de Origen de Datos

Luego de establecida la conexión con el origen de datos, se seleccionan las tablas que van a formar parte del Datamart.

Después de definir la conexión a un origen de datos, se crea una vista que identifica los datos concretos pertinentes para el modelo a usar.

La vista del origen de datos también permite personalizar la manera en que los datos del origen de datos se proporcionan al modelo de minería de datos.

En la Figura 6.3 se muestra el diseño de la vista del origen de datos.

En la Figura 6.4 se muestra el diagrama del diseño de la vista del origen de datos.

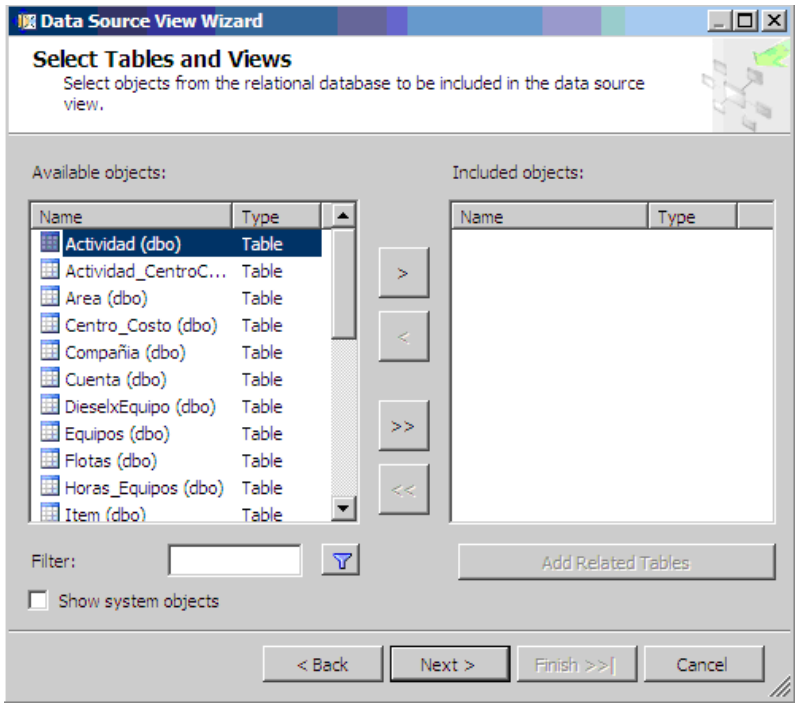


Figura 6.3 Diseño de la Vista del Origen de Datos

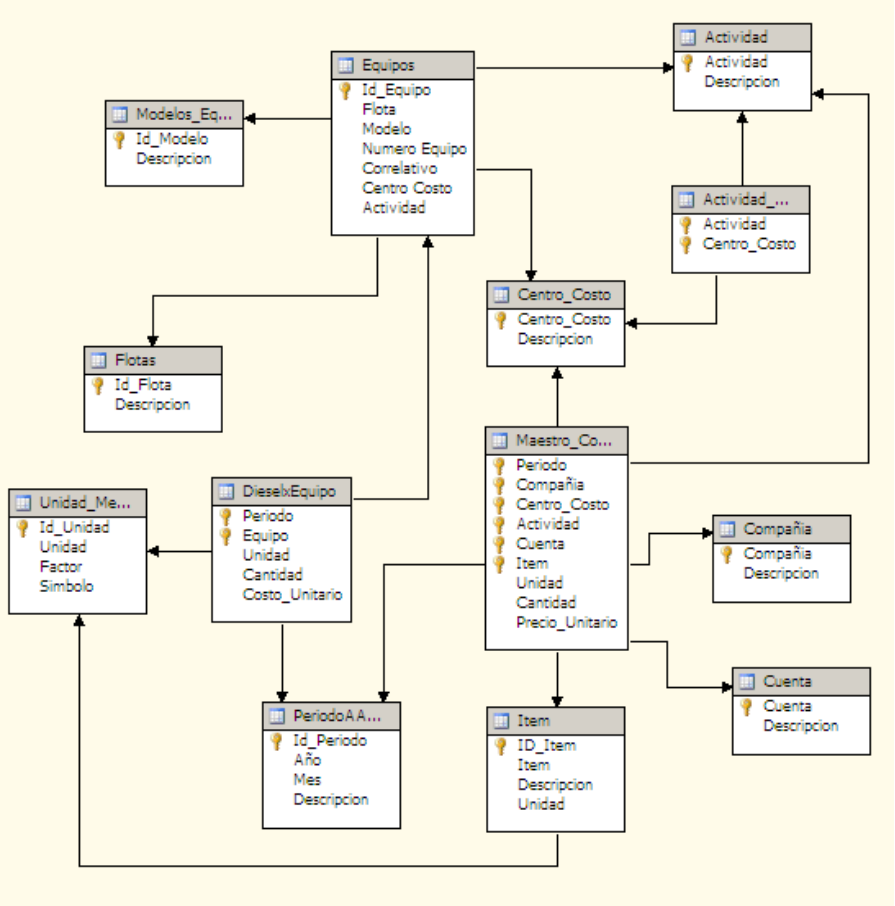


Figura 6.4 Diagrama del diseño de la Vista del Origen de Datos

6.2.3 Creación del Cubo "Presupuestos"

En esta etapa diseñamos el cubo "Presupuestos", donde vamos a definir las dimensiones, medidas y jerarquías del mismo.

El primer paso es la selección de la tabla "Maestro_Costos" que contiene las medidas (Figura 6.5); luego, se eligen las medidas a tomar en cuenta para crear el cubo (Figura 6.6), se seleccionan las dimensiones que tendrá el cubo (Figura 6.7); por último, se nombra y revisa la estructura del cubo "Presupuestos" (Figura 6.8).

Una vez creado el cubo, se revisa la vista del origen de datos del cubo "Presupuestos", como se observa en la Figura 6.9.

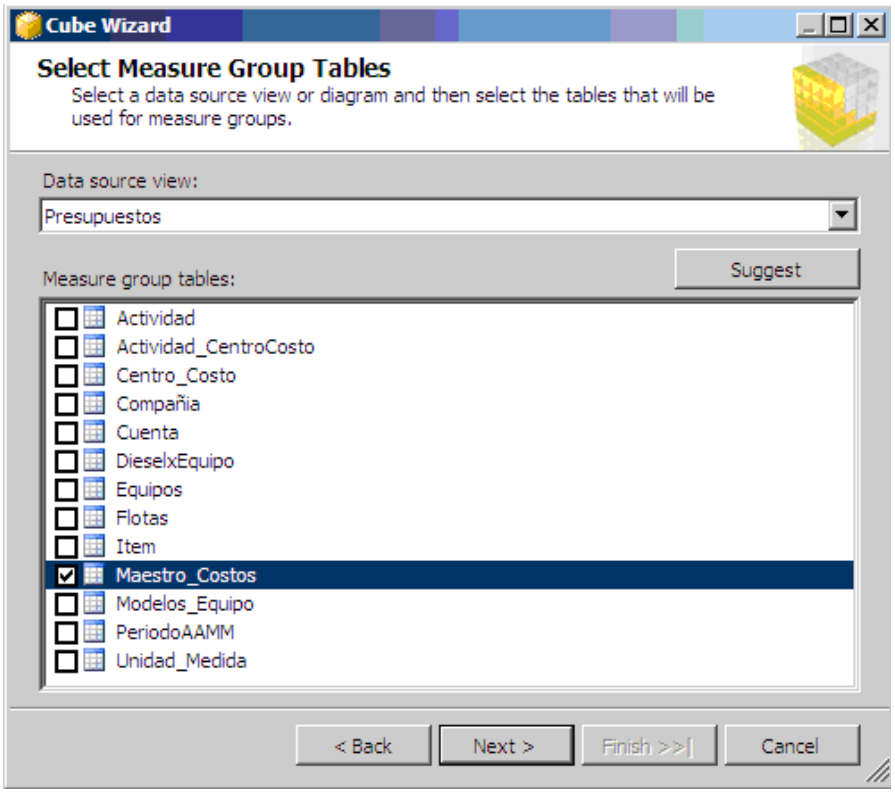


Figura 6.5 Selección de la tabla que contiene las medidas (valores numéricos)

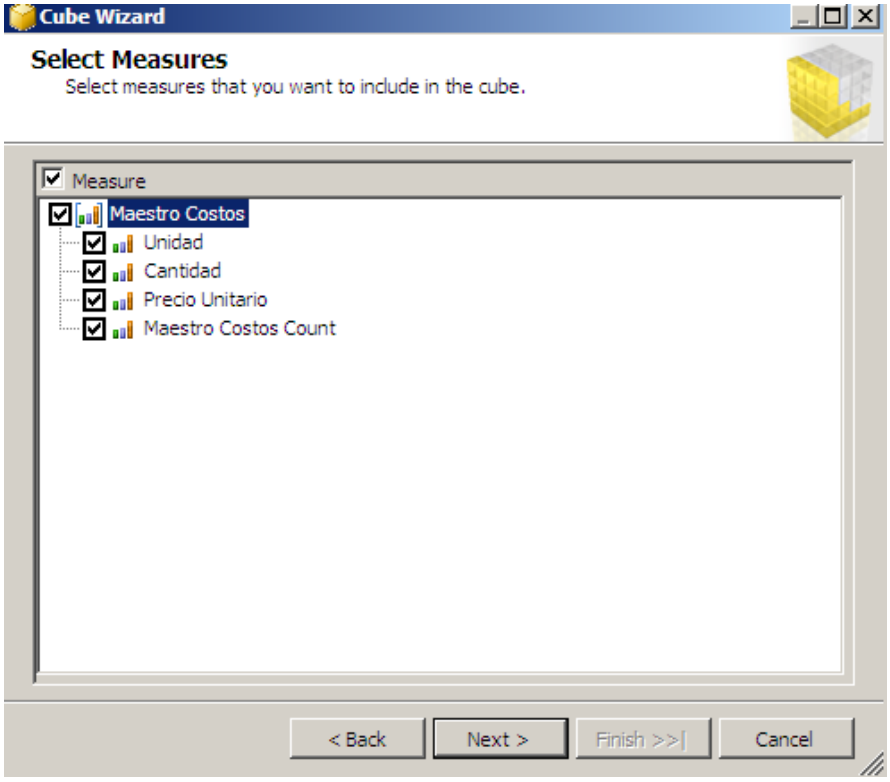


Figura 6.6 Selección de las medidas del cubo

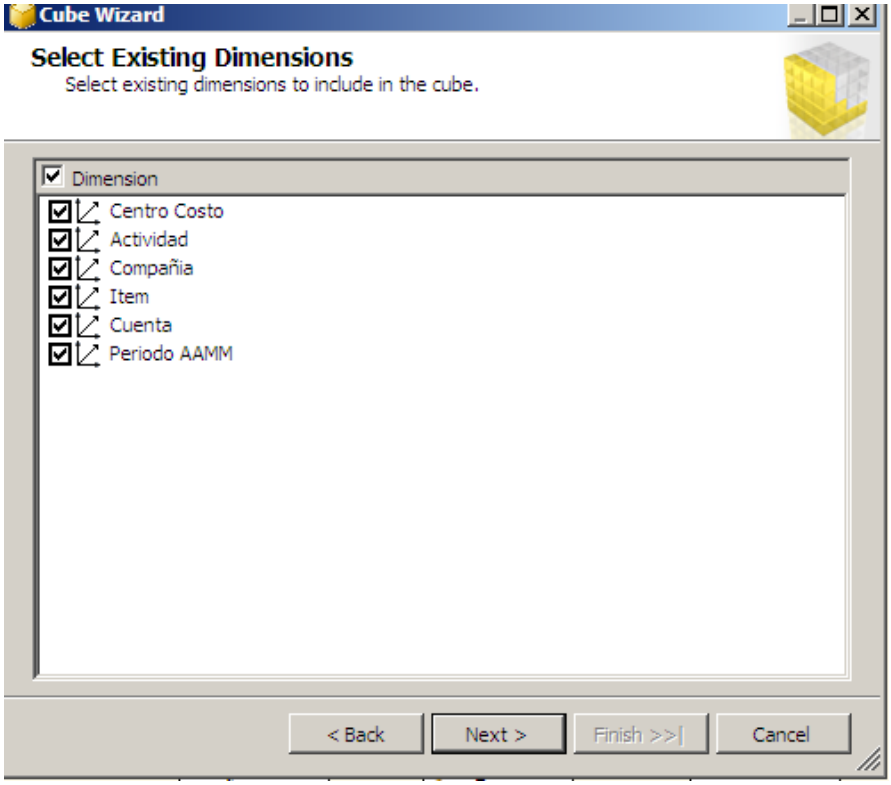


Figura 6.7 Selección de las dimensiones del cubo

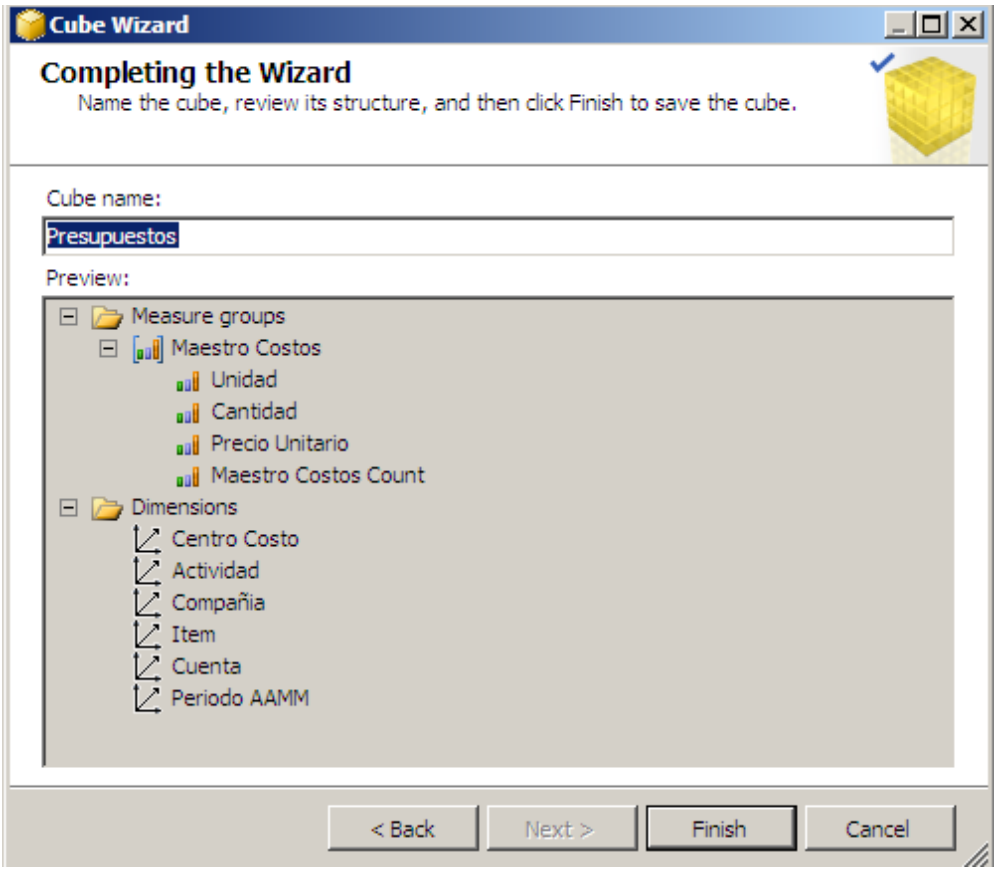


Figura 6.8 Creación del cubo Presupuestos

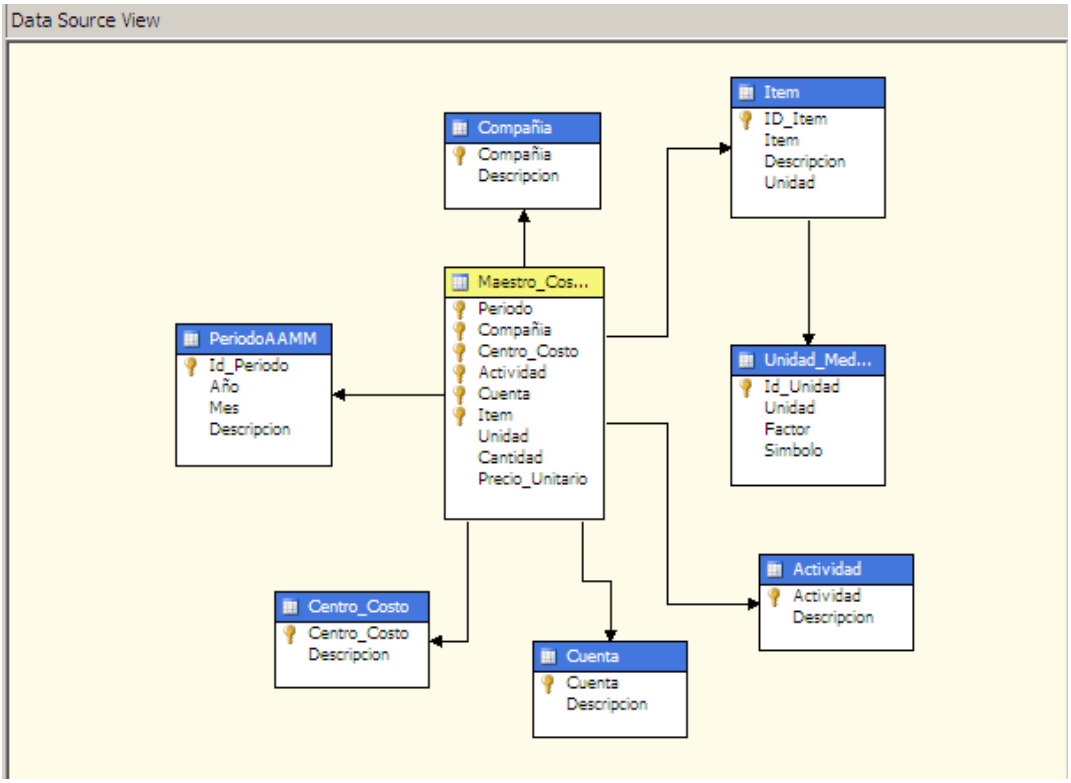


Figura 6.9 Vista de Origen de Datos del Cubo Presupuestos

Luego de crear el cubo, éste se procesa y queda listo para ser consultado desde cualquier aplicación haciendo uso de código MDX.

Los datos que residen en el cubo ofrecen información histórica de manera estructurada, la cual es usada en nuestra solución de minería de datos y que detallaremos más adelante.

6.3 Diseño del modelo de Minería de Datos

Cada proyecto de minería de datos contiene cuatro tipos de objetos:

- Orígenes de datos.
- Vistas del Origen de Datos, los cuales se basan en los orígenes de datos.
- Estructuras de Minería de Datos, que definen cómo se utilizan los datos en el modelo.
- Modelos de Minería de Datos, que crean y almacenan los patrones.

Los dos primeros objetos han sido definidos en el apartado anterior (6.2 Diseño del Datamart “Presupuestos”), por lo que, en adelante, nos centraremos en la definición de la Estructura y el Modelo de Minería de Datos.

6.3.1 Estructura de Minería de Datos

Una estructura de minería de datos define las columnas de datos y las columnas de tablas anidadas, que se obtienen de la vista del origen de datos o de un cubo OLAP en el proyecto.

Para definir la estructura de minería de datos nos basamos en el cubo “Presupuestos” (Figura 6.10); luego, seleccionamos a las series temporales como la técnica a emplear para el modelo de minería de datos (Figura 6.11), elegimos el recurso “Periodo AAMM” que será la dimensión del cubo (Figura 6.12), como atributo clave para el modelo se selecciona “IdPeriodo” (Figura 6.13), se especifican las columnas a ser usadas en la estructura de minería de datos (Figura 6.14), se definen si las columnas son de “Entrada” o de “Predicción” (Figura 6.15), se eligen los tipos de datos de los atributos (Figura 6.16), finalmente, se nombra a la estructura de minería de datos como “Periodo AAMM” (Figura 6.17) y se visualiza esta creación en la Figura 6.18.

En la Figura 6.19 podemos observar el procesamiento del proyecto, basado en el datamart “Presupuestos” y la estructura de minería de datos “Periodo AAMM”.

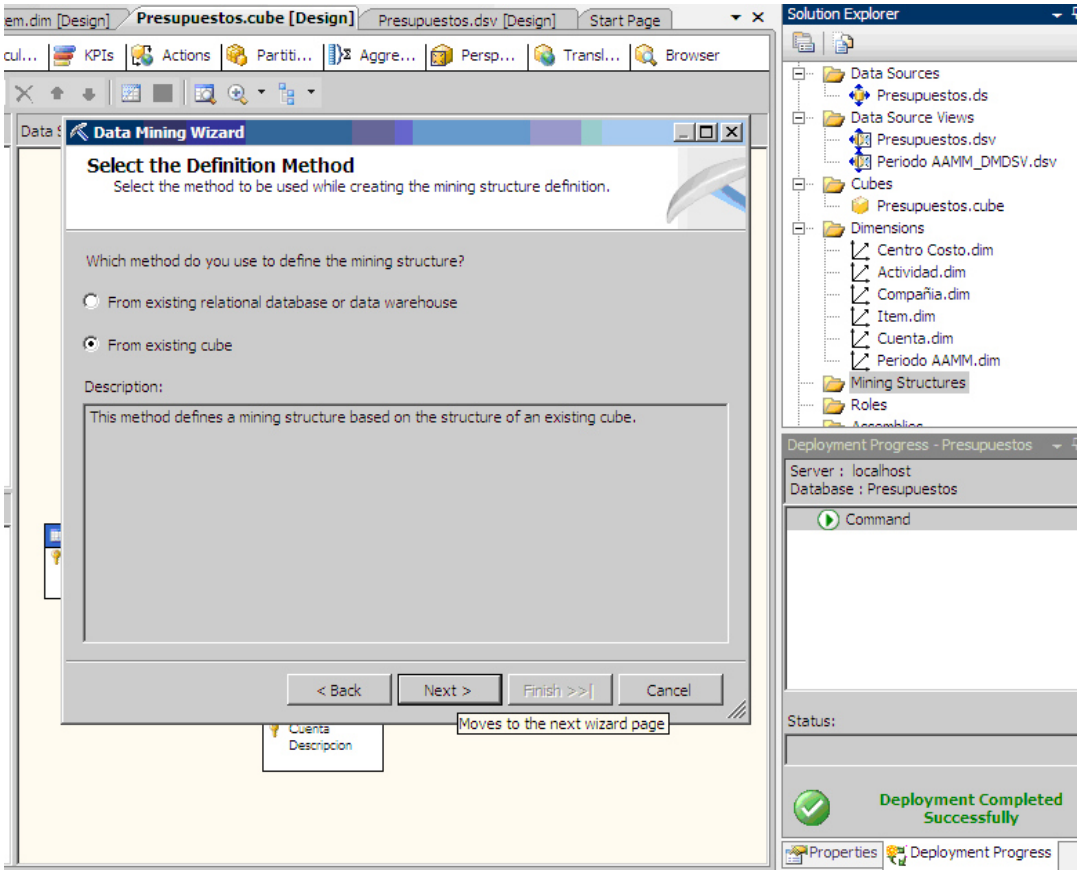


Figura 6.10 Definición de la Estructura de Minería de Datos basado en el Cubo Presupuestos

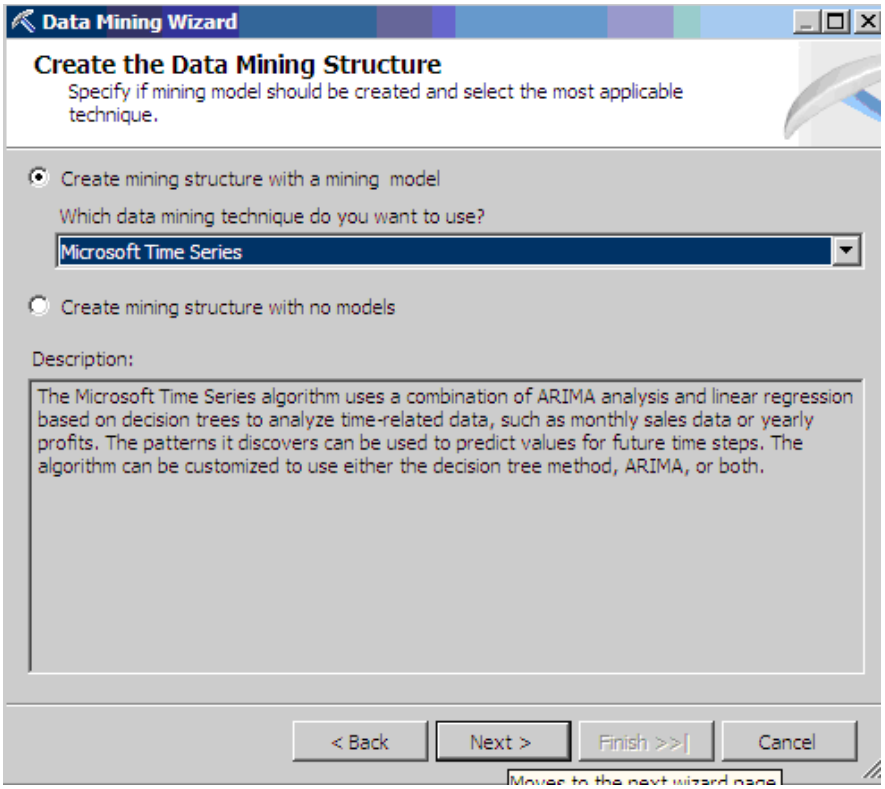


Figura 6.11 Selección de la Técnica a emplear para el modelo de minería de datos

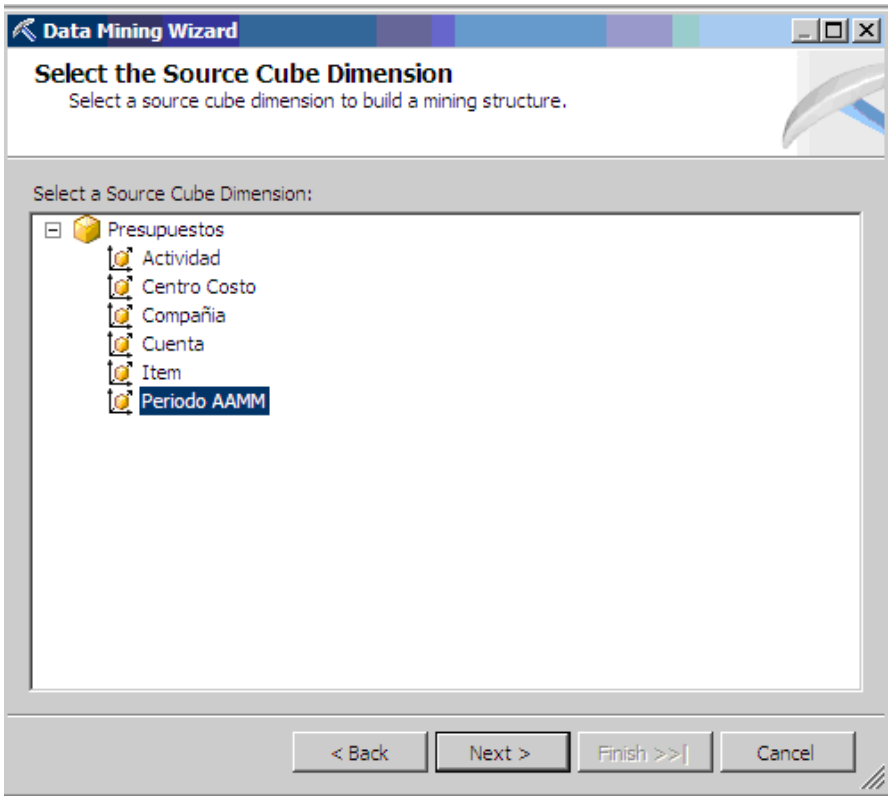


Figura 6.12 Selección del recurso de la dimensión del cubo

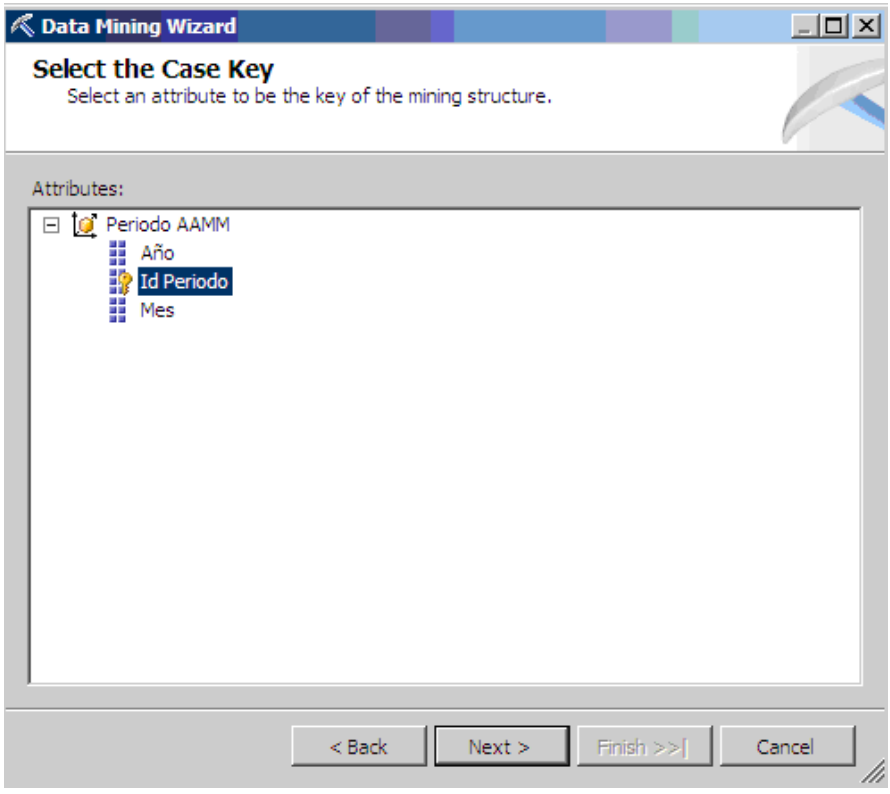


Figura 6.13 Selección del atributo clave para el modelo

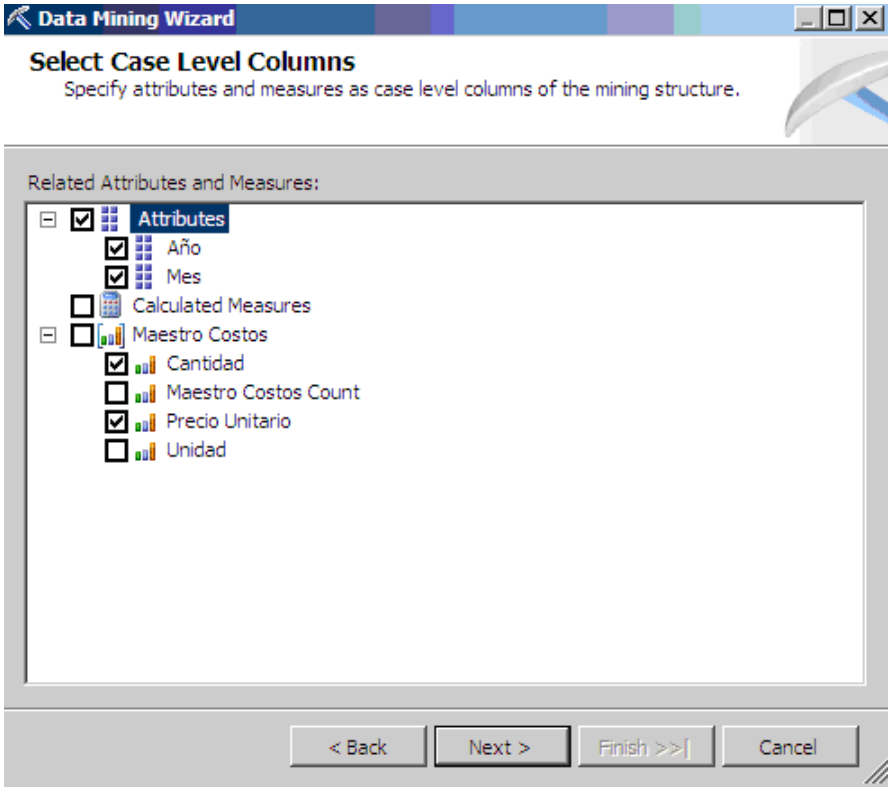


Figura 6.14 Selección de las columnas usadas en la Estructura de minería de datos

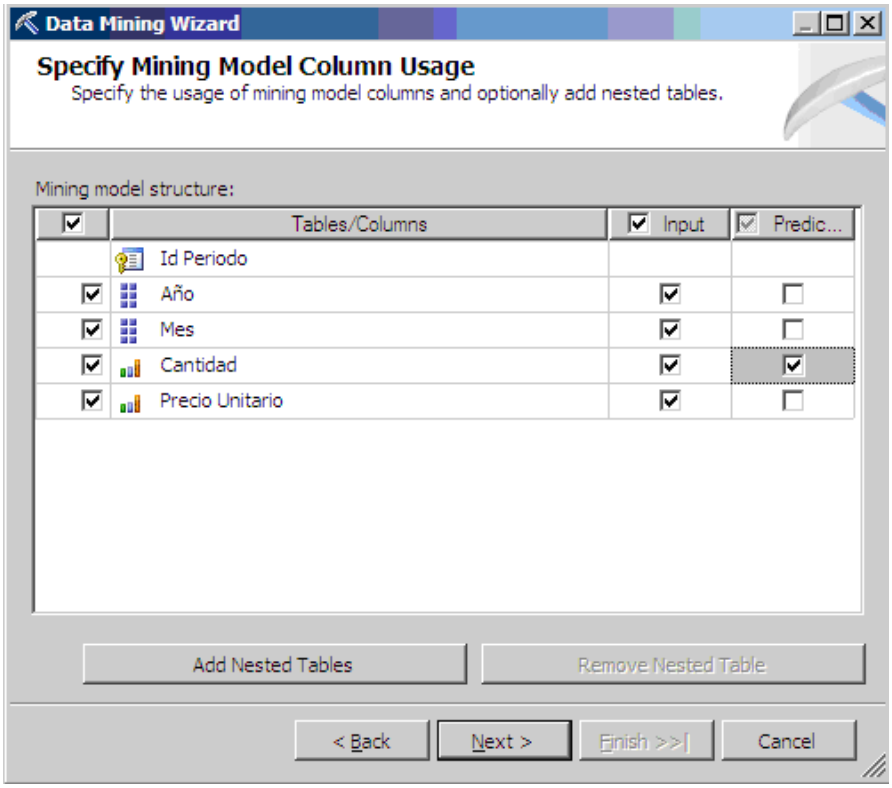


Figura 6.15 Definición de los tipos de columnas: Entrada o Predicción

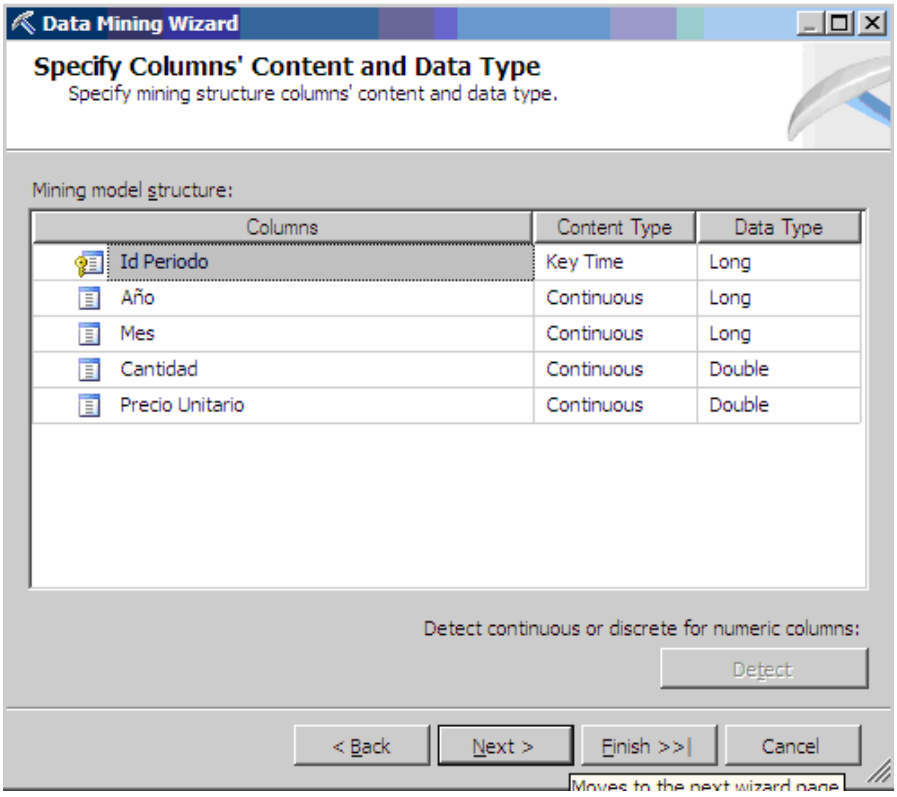


Figura 6.16 Definición del tipo de datos de los atributos

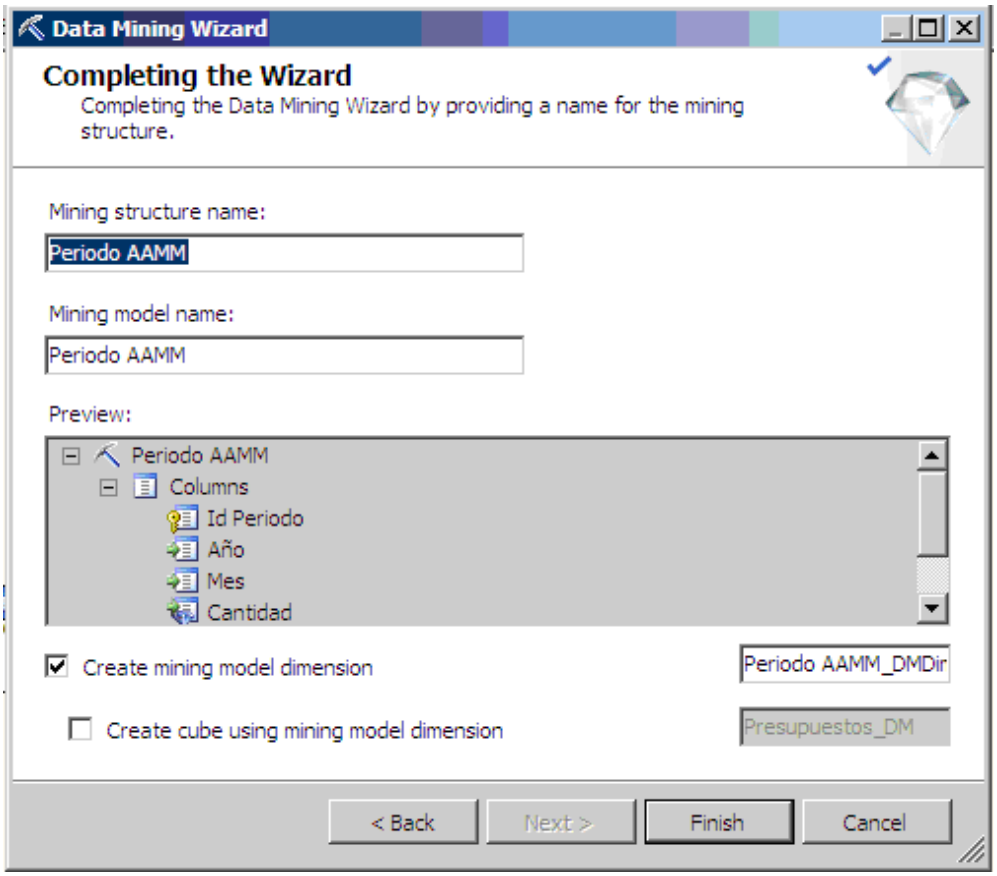


Figura 6.17 Ingreso de datos finales para la generación de la estructura de minería de datos

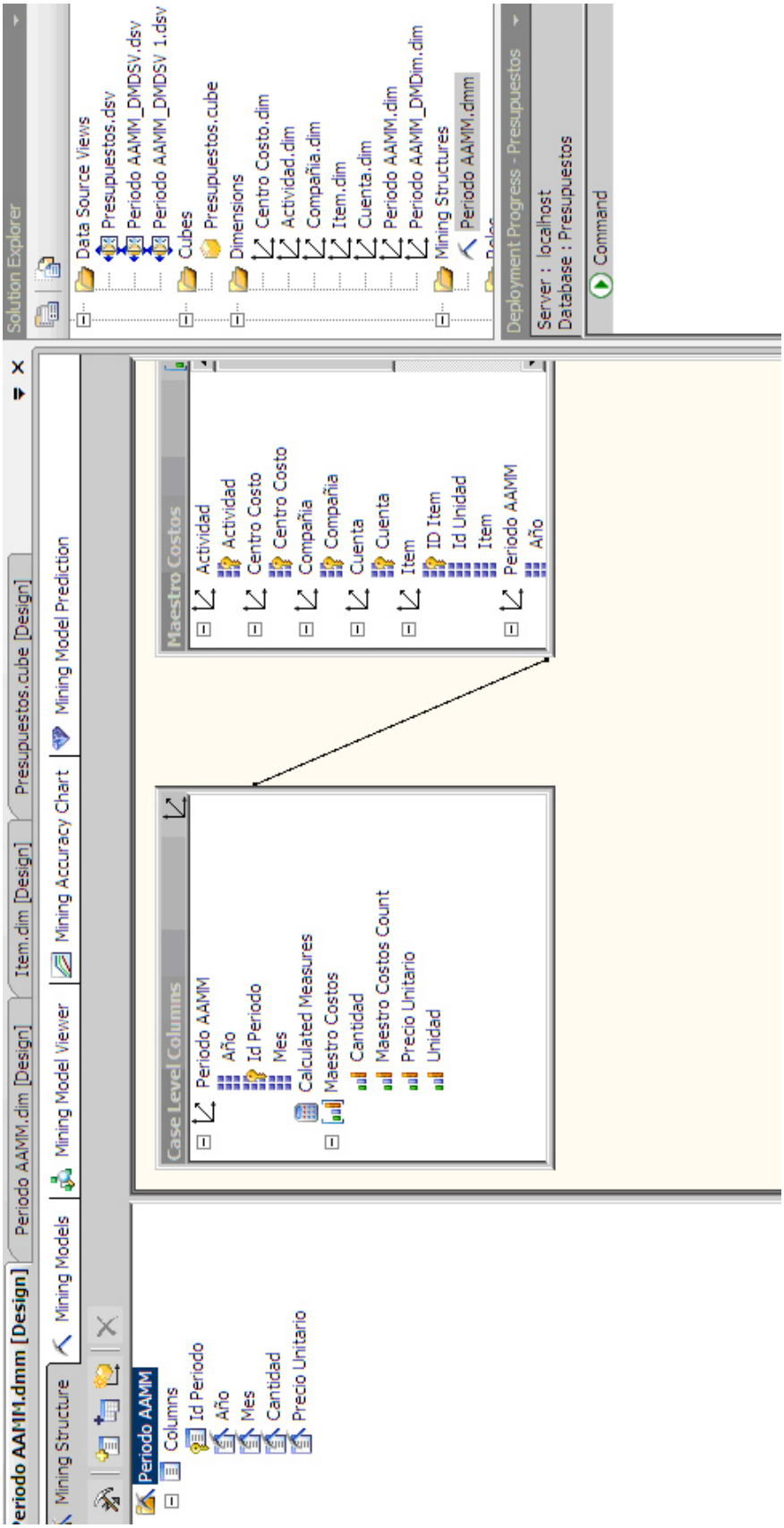


Figura 6.18 Estructura de minería de datos

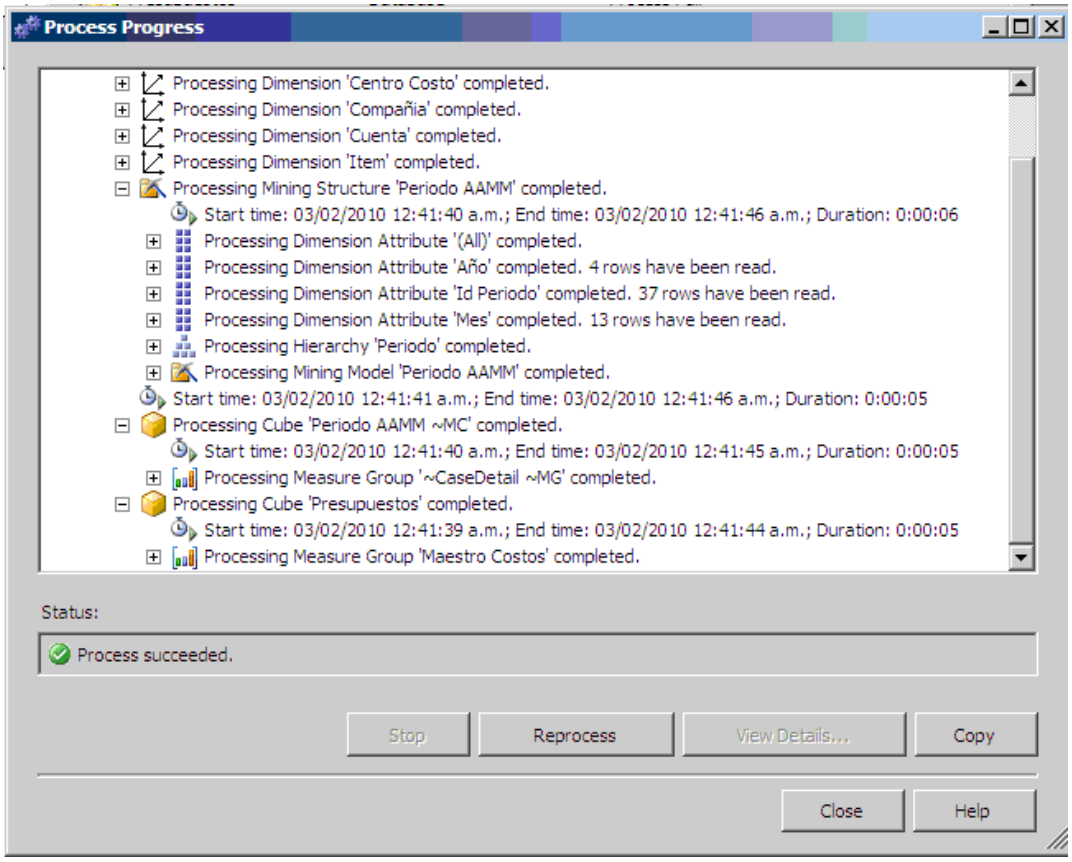


Figura 6.19 Procesamiento del Proyecto: Datamart de Presupuestos y Estructura de Minería de Datos

6.3.2 Modelo de Minería de Datos

En el modelo de minería de datos definimos el algoritmo o método de análisis que se utilizará en los datos.

Se procesa el modelo ejecutando los datos en la vista del origen de datos a través del algoritmo elegido: Series Temporales, el cual genera un modelo matemático de los datos.

Posteriormente, exploramos visualmente el modelo de minería de datos (Figura 6.20) y creamos las consultas de predicción del mismo.

La herramienta de Analysis Services proporciona varias opciones para procesar los objetos del modelo de minería de datos, incluyendo la capacidad de controlar cuáles y cómo se procesan los objetos.

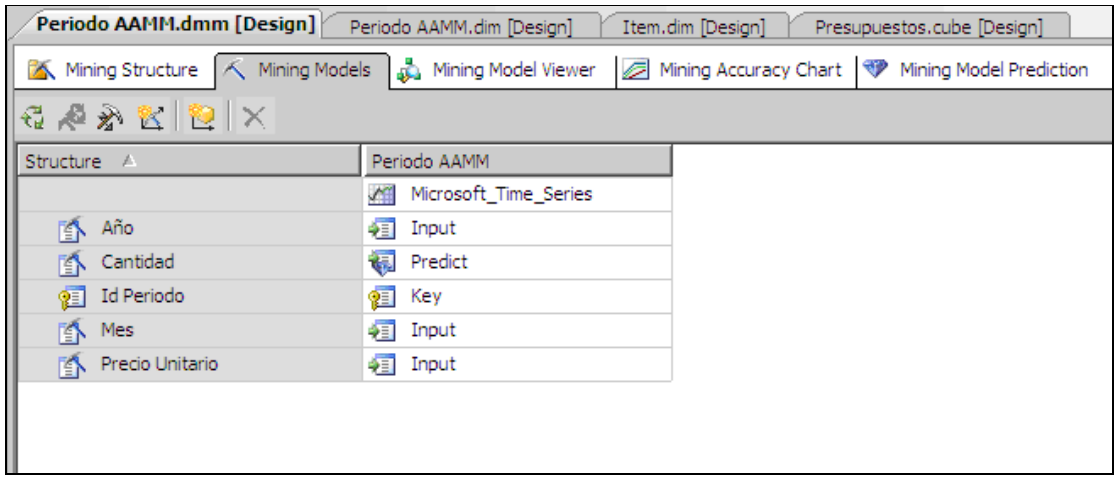


Figura 6.20 Visualización de los valores de entrada y de predicción del modelo

Después de crear el modelo, investigamos los resultados y comprobamos que el modelo de Series Temporales se comporta mejor en el tiempo.

Esto lo podemos apreciar en la ficha Visor de modelos de minería de datos en el Diseñador de minería de datos del Analysis Services (Figuras 6.21 y 6.22), el cual proporciona visores para cada tipo de modelo de minería de datos, los que nos ayudaron a explorar los modelos.

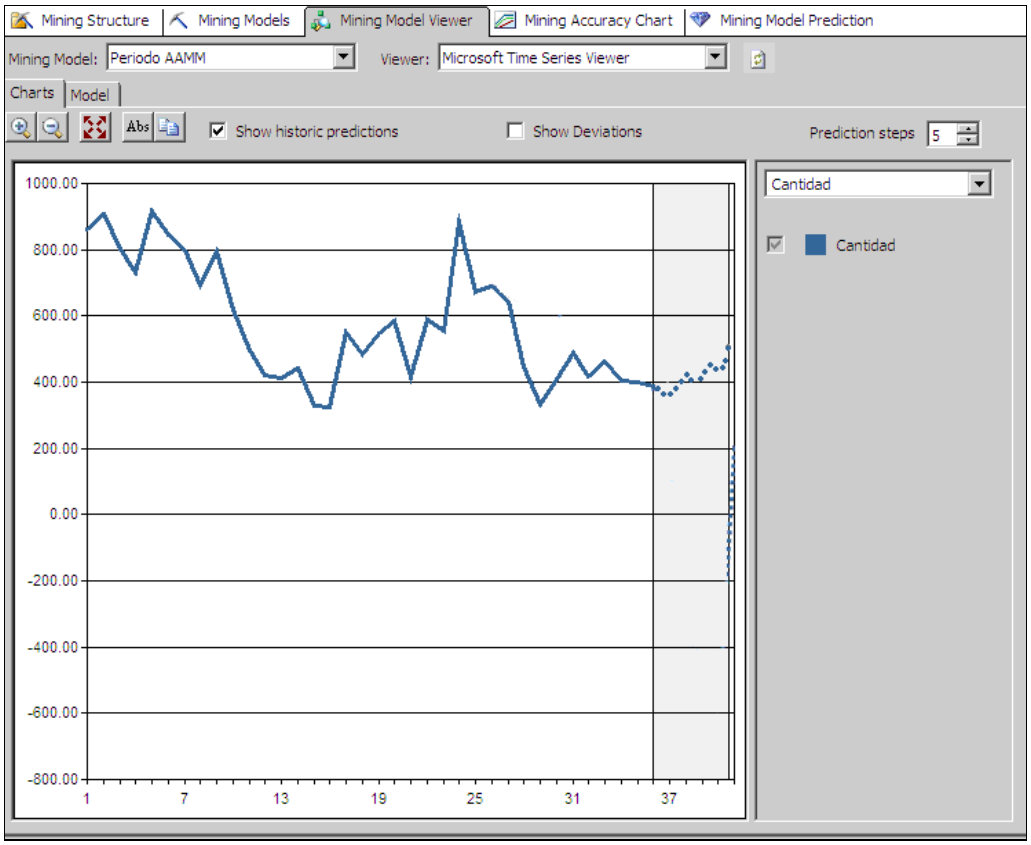


Figura 6.21 Vista gráfica del modelo de minería de datos en el Visor de modelos

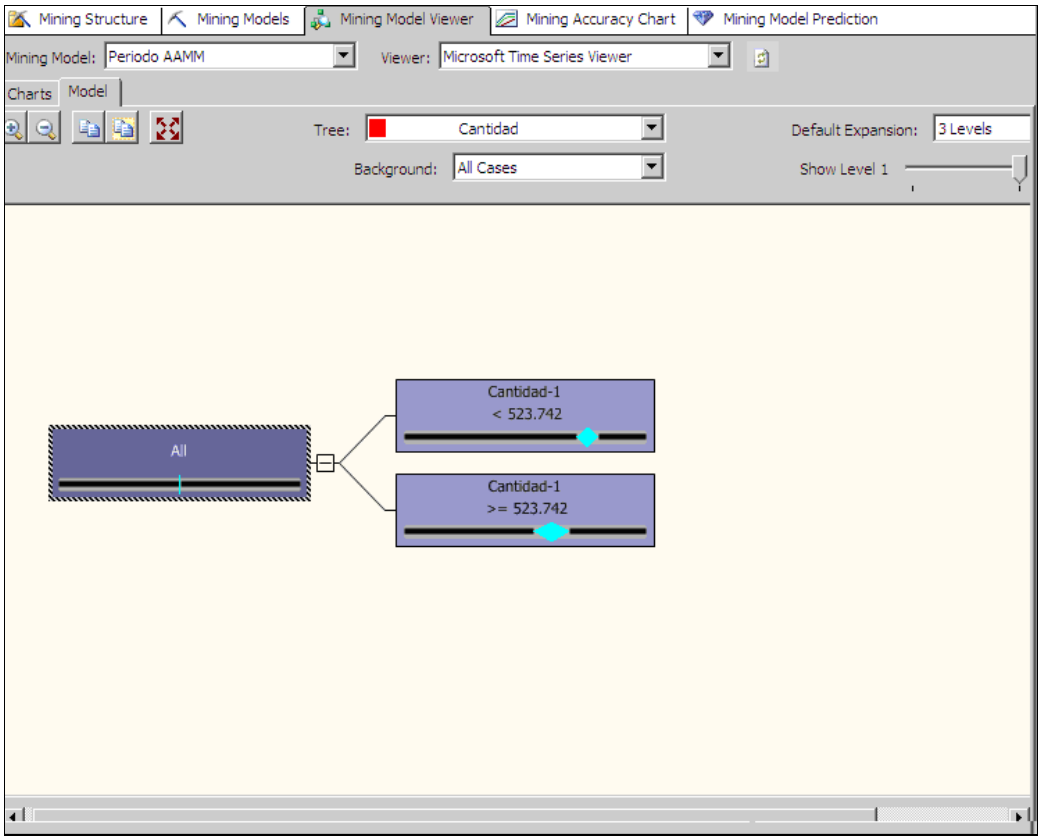


Figura 6.22 Vista del modelo de minería de datos en el visor de modelos

Se usó el informe de validación cruzada, nuevo atributo en SQL Server 2008, para realizar un submuestreo reiterativo de los datos y determinar si el modelo escogido se inclina a un conjunto determinado de datos. Las estadísticas que este informe proporciona se utilizaron para comparar objetivamente los modelos y evaluar la calidad de los datos de prueba.

En la Figura 6.23 se muestra el modelo de minería de datos obtenido, el cual puede ser usado con una nueva tabla de entrada de datos. En nuestro caso, esto no fue necesario porque contamos con una sola tabla de entrada de datos base.

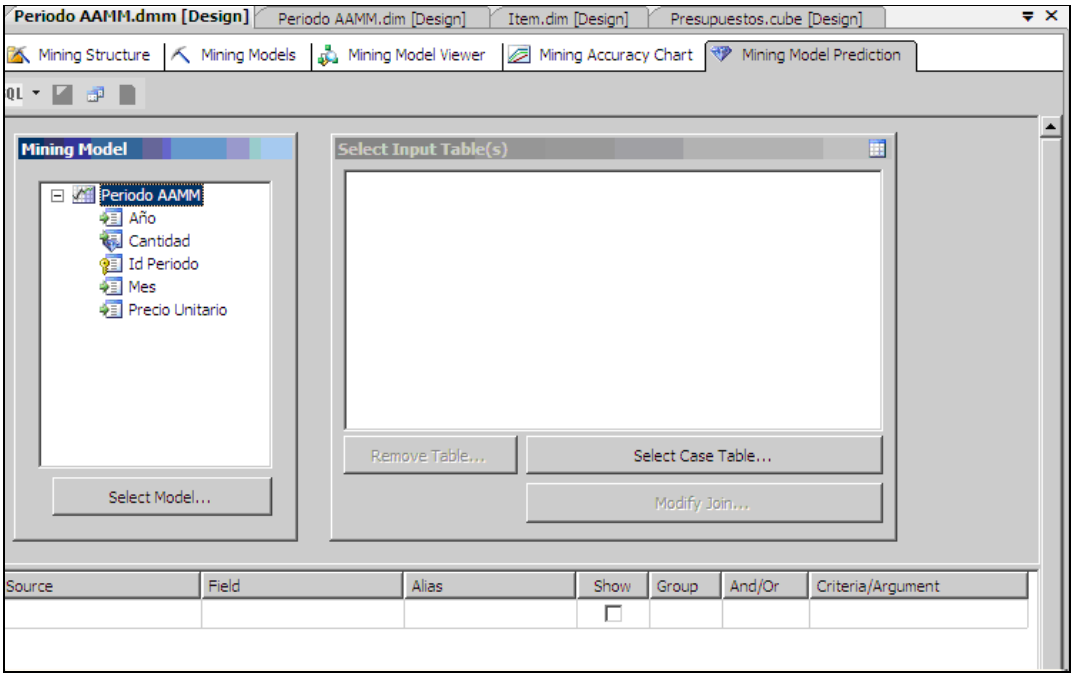


Figura 6.23 Generación de modelos predictivos a partir de una nueva tabla de entrada de datos

Después de explorar y comparar los modelos de minería de datos, utilizamos la herramienta de creación de predicciones de Analysis Services que ofrece un lenguaje de consulta denominado Extensiones de Minería de Datos (DMX – Data Mining eXtensions), que es la base para la creación de predicciones y es fácilmente convertible en script.

Este lenguaje ha sido insertado en nuestra aplicación para mostrar las predicciones realizadas por el modelo de minería de datos.

6.4 Modelado del Negocio

Para el modelado del negocio se han descrito el actor del negocio, los casos de uso del negocio (CUN), los trabajadores del negocio y el diagrama de actividades de cada CUN.

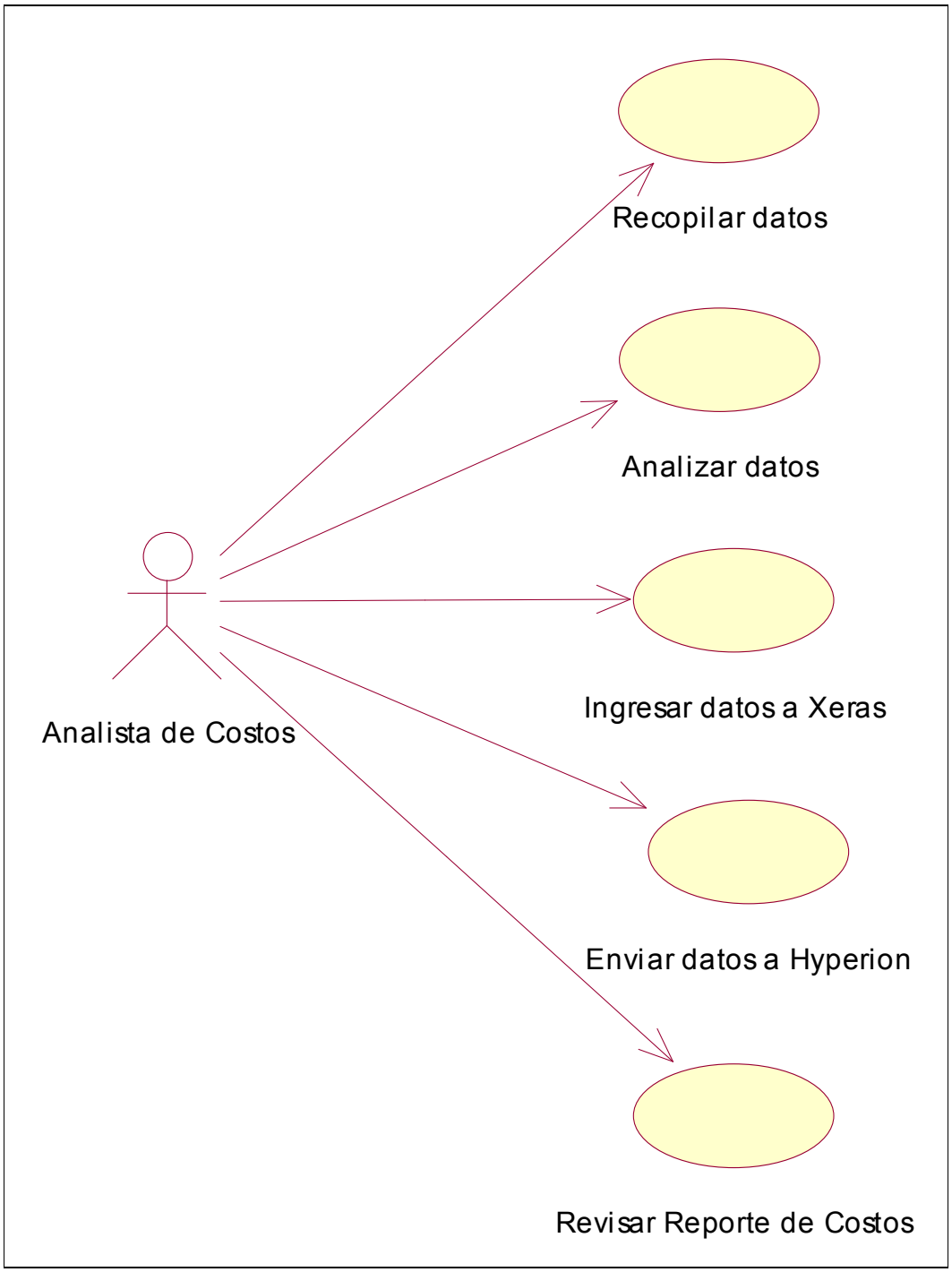
6.4.1 Listado de Actores

Actor del negocio	Descripción
1. Analista de Costos	Persona que elabora el presupuesto del área de Operaciones Mina. Para ello realiza los procesos de recopilación de datos de diversas fuentes, análisis de datos para la obtención de ratios de consumos, ingreso de datos a Xeras, envío de datos a Hyperion y finalmente revisión del Reporte de Costos.

6.4.2 Descripción de los Casos de Uso del Negocio (CUN)

Casos de uso del negocio	Descripción
1. CUN01: Recopilar datos	El caso de uso comienza cuando el Analista de costos busca los datos de producción y consumos que se encuentran en diversas fuentes de datos. El caso de uso termina cuando el Analista de costos consolida todos estos datos en un solo archivo.
2. CUN02: Analizar datos	El caso de uso comienza cuando el Analista de costos realiza cálculos estadísticos sobre los datos recopilados. El caso de uso termina cuando se obtienen los ratios de los consumibles.
3. CUN03: Ingresar datos a Xeras	El caso de uso comienza cuando el Analista de costos ingresa los ratios de consumibles, entre otros datos (precios, horas requeridas, plan de minado) al sistema Xeras. El caso de uso termina cuando el sistema Xeras genera el reporte de cantidades, precios unitarios y costos totales.
4. CUN04: Enviar datos a Hyperion	El caso de uso comienza cuando el Analista de Costos envía el reporte de Xeras al Administrador de Hyperion. El caso de uso termina cuando el Administrador de Hyperion ha subido los datos del reporte de Xeras a Hyperion.
5. CUN05: Revisar Reporte de Costos	El caso de uso comienza cuando el Analista de costos revisa los datos que se han subido al Hyperion. El caso de uso termina cuando el Analista de costos aprueba los datos almacenados en Hyperion.

Diagrama de Casos de Uso del Negocio



6.4.3 Lista de Trabajadores del Negocio

Trabajador del negocio	Descripción
1. Administrador de Hyperion	Encargado de recibir el reporte del sistema Xeras con los datos de consumo, precios unitarios y costos, para cargarlos al sistema Hyperion, ejecutar el proceso de actualización de datos e informar al Analista de Costos sobre la culminación de este proceso.
2. Supervisor de Planeamiento a Largo Plazo	Supervisor del área de Ingeniería que está encargado de elaborar el plan de minado para enviárselo al Analista de Costos, quien se encargará de costearlo. También se encarga de mantener actualizados los datos de producción en las diversas fuentes establecidas para ello (hojas de cálculo, bases de datos).
3. Sistema Xeras	Sistema que calcula los consumos y costos finales en base a los datos ingresados por el Analista de Costos.
4. Sistema Hyperion	Sistema que almacena los consumos, precios unitarios y costos finales para que estén accesibles por cualquier usuario de este sistema.

6.4.4 Especificación de los Casos de Uso del Negocio

A continuación se describe cada uno de los casos de uso del negocio.

CUN01: Recopilar Datos

Sistema: Sistema de Proyección de Costos	
Nombre del Caso de Uso	Recopilar Datos
Código del Caso de Uso	CUN01
Actores participantes	Analista de Costos
Descripción	El caso de uso comienza cuando el Analista de costos busca los datos de producción y consumos que se encuentran en diversas fuentes de datos. El caso de uso termina cuando el Analista de costos consolida todos estos datos en un solo archivo.
Condición Inicial	El gasto de todos los consumibles se registra en una base de datos de Oracle de manera automática.
Flujo Básico de Eventos	1. El analista de costos verifica que los datos estén almacenados en las diversas fuentes de datos establecidas, como son la base de datos de horas operativas de equipos (Powerview en Sistema Dispatch), la base de datos de consumo de diesel (Consumos en Microsoft Access), y el plan de producción en hojas de cálculo, principalmente. 2. El analista de costos consolida todos los datos de consumos y horas operativas en un solo archivo. 3. El analista de costos guarda este archivo en una carpeta de costos ubicada en el servidor \\perhuafs1.
Flujos Alternativos	1.1.A Si los datos de producción y horas de equipos requeridos no se encuentran almacenados, el analista de costos solicita al supervisor de Ingeniería que registre estos datos manualmente. 1.1.B Si los datos de consumo de diesel no se encuentran actualizados en la base de datos de Primax (Microsoft Access), solicita al Supervisor de Primax que haga la carga de los consumos de diesel en su base de datos.

CUN02: Analizar Datos

Sistema: Sistema de Proyección de Costos	
Nombre del Caso de Uso	Analizar Datos
Código del Caso de Uso	CUN02
Actores participantes	Analista de Costos
Descripción	El caso de uso comienza cuando el Analista de costos realiza cálculos estadísticos sobre los datos recopilados. El caso de uso termina cuando se obtienen los ratios de los consumibles.
Condición Inicial	Los datos deben estar clasificados por tipo de consumible y pertenecer a un período de tiempo no menor a 12 meses.
Flujo Básico de Eventos	1. El analista de costos realiza un análisis estadístico sobre los datos recopilados. 2. El analista de costos obtiene ratios de consumos de los principales consumibles, como son: diesel (gal/hr), emulsión matriz (factor de carga), nitrato de amonio (factor de carga).
Flujos Alternativos	No se consideran flujos alternativos.

CUN03: Ingresar datos a Xeras

Sistema: Sistema de Proyección de Costos	
Nombre del Caso de Uso	Ingresar datos a Xeras
Código del Caso de Uso	CUN03
Actores participantes	Analista de Costos
Descripción	El caso de uso comienza cuando el analista de costos y el supervisor de planeamiento a largo plazo ingresan los ratios de consumibles, entre otros datos (precios, horas requeridas, plan de minado) al sistema Xeras. El caso de uso termina cuando se genera una hoja de cálculo en el sistema Xeras, el cual es un reporte de cantidades, precios unitarios y costos totales.
Condición Inicial	Se deben tener configurados el escenario sobre el cual se trabajará en Xeras y los recursos físicos, como el calendario, la cantidad de equipos disponibles, el rol de trabajo del personal operativo, entre otros.
Flujo Básico de Eventos	1. El analista de costos ingresa los ratios de consumos de diesel, nitrato de amonio y emulsión matriz en el sistema Xeras. 2. El analista de costos ingresa los precios unitarios de los consumibles en el sistema Xeras. 3. El supervisor de planeamiento a largo plazo ingresa el plan de minado al sistema Xeras. 4. El supervisor de planeamiento a largo plazo ingresa las horas requeridas de equipos para cumplir el plan de producción, en el sistema Xeras. 5. El analista de costos ejecuta el sistema Xeras para obtener las cantidades y costos finales. 6. El analista de costos ejecuta, en el sistema Xeras, la opción de Generar Reporte de Hyperion, con el cual se obtiene de cantidades, precios unitarios y costos totales en una hoja de cálculo.
Flujos Alternativos	5.1 Si los datos obtenidos no son coherentes con lo que el analista de costos desea usar, entonces el analista de costos puede ajustar los ratios de consumibles, de acuerdo a su experiencia.

CUN04: Enviar datos a Hyperion

Sistema: Sistema de Proyección de Costos	
Nombre del Caso de Uso	Enviar datos a Hyperion
Código del Caso de Uso	CUN04
Actores participantes	Analista de Costos
Descripción	El caso de uso comienza cuando el Analista de Costos envía el reporte de Xeras al Administrador de Hyperion. El caso de uso termina cuando el Administrador de Hyperion ha subido los datos del reporte de Xeras a Hyperion.
Condición Inicial	El administrador de Hyperion debe haber creado la versión working en Hyperion para poder cargar los datos, como también el escenario del presupuesto en curso (especificando si el presupuesto será Budget o Forecast y de qué año se trata, por ejemplo para nombrar al Forecast F10+2 del 2010 usará la denominación 2010Fcst10+2).
Flujo Básico de Eventos	1. El analista de costos envía al administrador de Hyperion el reporte generado en Xeras. 2. El administrador de Hyperion carga este reporte de cantidades, precios unitarios y costos totales al Hyperion, en la versión y escenario creados para dicho presupuesto. 3. El administrador de Hyperion ejecuta el proceso de actualización de datos en Hyperion. 4. El administrador de Hyperion informa al analista de costos la correcta carga de los datos en Hyperion.
Flujos Alternativos	No se consideran flujos alternativos.

CUN05: Revisar Reporte de Costos

Sistema: Sistema de Proyección de Costos	
Nombre del Caso de Uso	Revisar Reporte de Costos
Código del Caso de Uso	CUN05
Actores participantes	Analista de Costos
Descripción	El caso de uso comienza cuando el Analista de costos revisa los datos que se han subido al Hyperion. El caso de uso termina cuando el Analista de costos aprueba los datos almacenados en Hyperion.
Condición Inicial	El analista de costos debe validar su usuario en el Sistema Hyperion.
Flujo Básico de Eventos	1. El analista de costos ingresa al formulario de costos del Sistema Hyperion. 2. El analista de costos verifica que los datos reflejados en el formulario de costos sean los mismos que los datos enviados en el Reporte de Hyperion generado en el sistema Xeras. 3. El analista de costos aprueba los datos almacenados en Hyperion.
Flujos Alternativos	2.1 Si los datos cargados en Hyperion no coinciden con los datos enviados, se realiza una nueva carga de los datos, regresando al CUN04: Enviar datos a Hyperion.

6.4.5 Diagramas de Actividades

A continuación se muestran los diagramas de actividades relacionados a cada caso de uso descrito.

Diagrama de Actividades del CUN01: Recopilar Datos

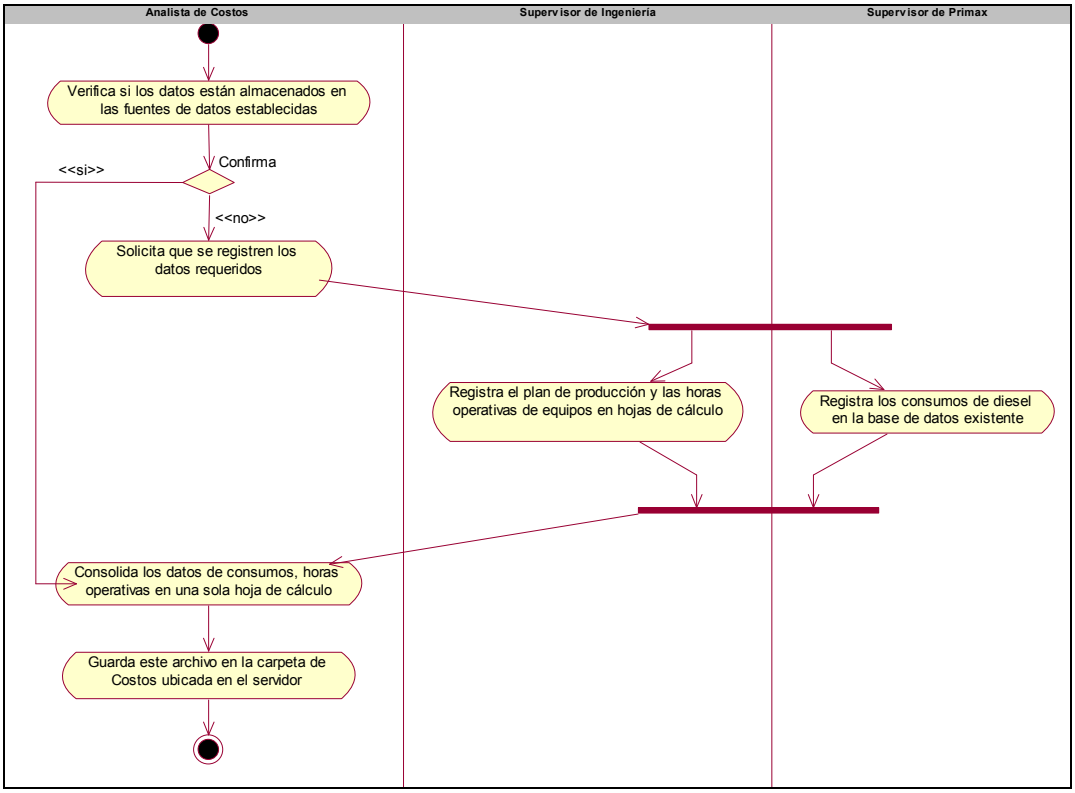


Diagrama de Actividades del CUN02: Analizar Datos

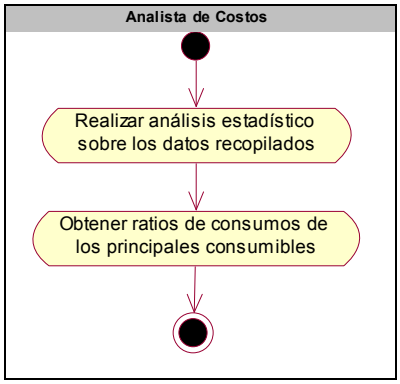


Diagrama de Actividades del CUN03: Ingresar datos a Xeras

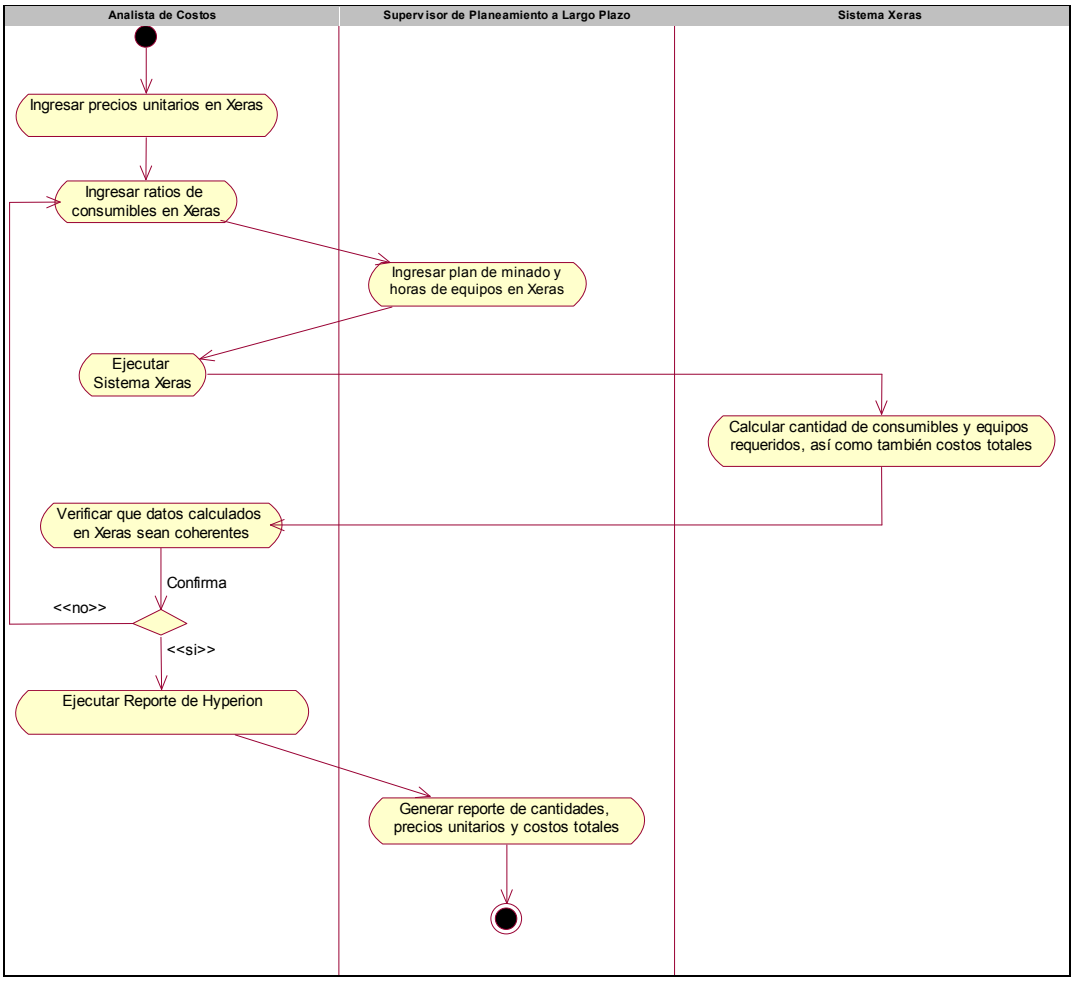


Diagrama de Actividades del CUN04: Enviar datos a Hyperion

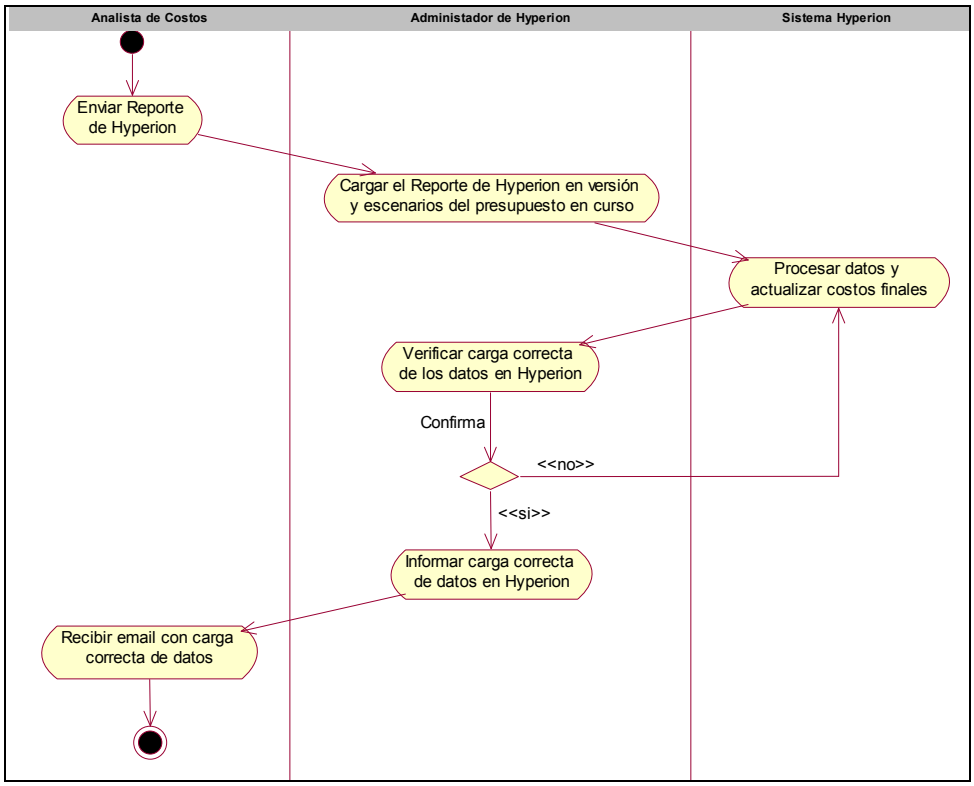
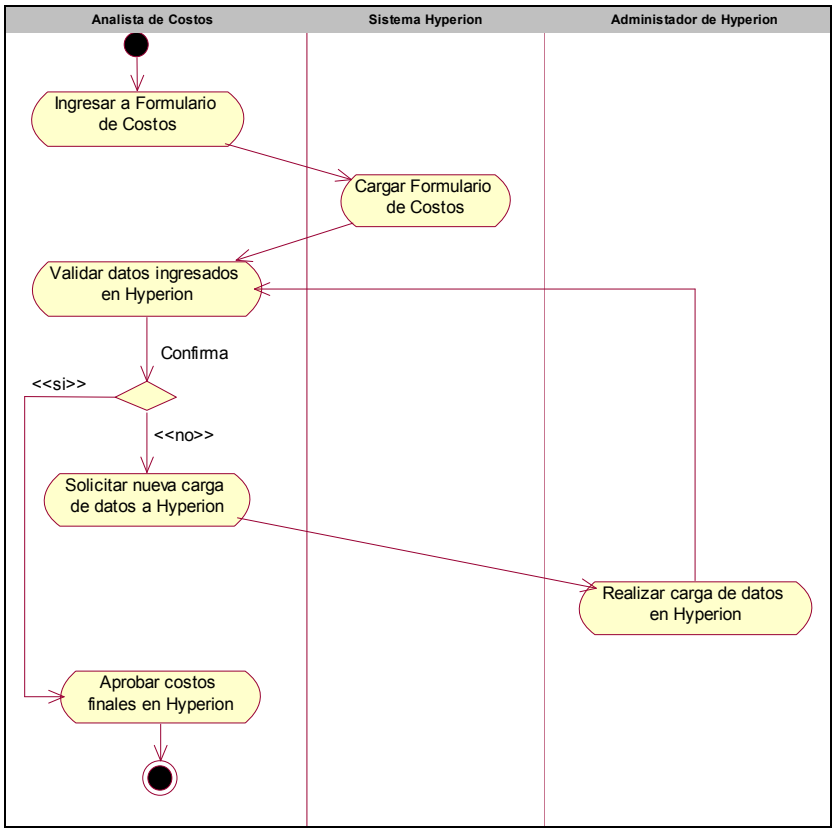


Diagrama de Actividades del CUN05: Revisar Reporte de Costos



6.5 Interface gráfica de la Aplicación

A continuación se muestra la interface gráfica de la implementación de la aplicación. En la Figura 6.24 se tiene la ventana de inicio para ingresar en el Sistema de Presupuestos.

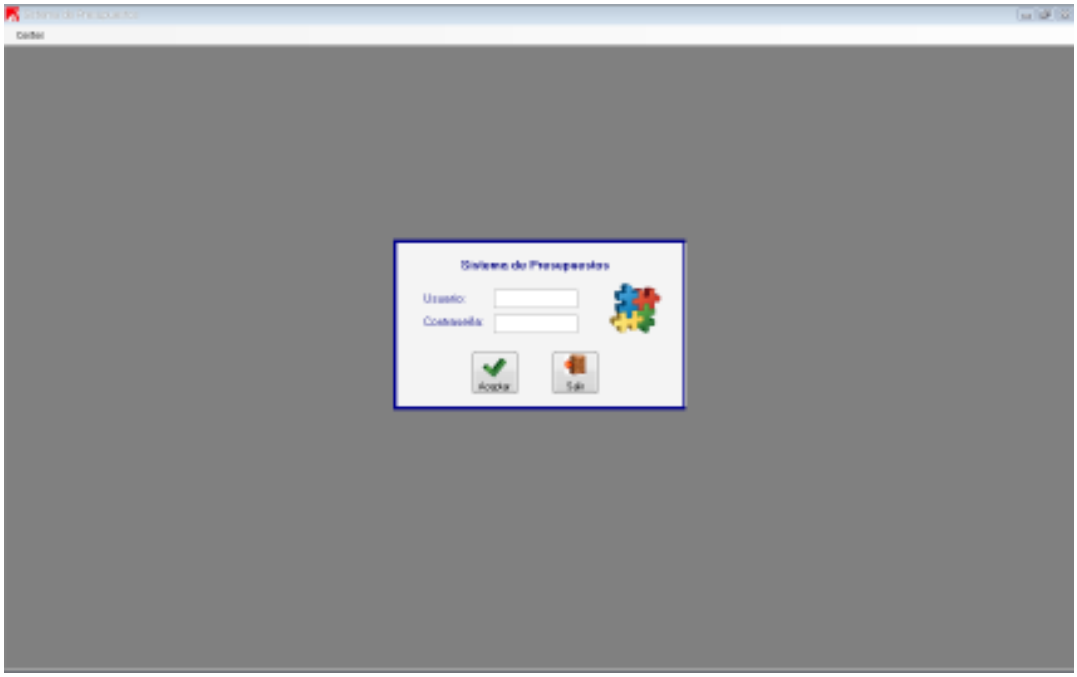


Figura 6.24 Ventana de inicio

En la Figura 6.25 se observa la ventana inicial del sistema de proyección de costos, donde se ingresan los parámetros de acuerdo a la información que se necesite obtener, dentro de los principales parámetros tenemos:

- Compañía: Por ser una multiempresa, tiene varias sedes (Compañía)
- Area: Es el área dentro de la organización (Operaciones, Mantenimiento, Procesos, etc.)
- Periodicidad: La cual indica si se va a generar un reporte a nivel de detalle mensual o anual.
- Periodo: Es el horizonte al cual se desea generar la información. Dependiendo de la periodicidad, puede ser a 3 meses, 12 meses, 1 año, 2 años, etc.
- Estructura: En esta sección elegimos a un mayor nivel de detalle en caso se desee procesar información a un determinado Centro de Costo, Actividad o Cuenta.

La Figura 6.26 muestra alguna de estas opciones.

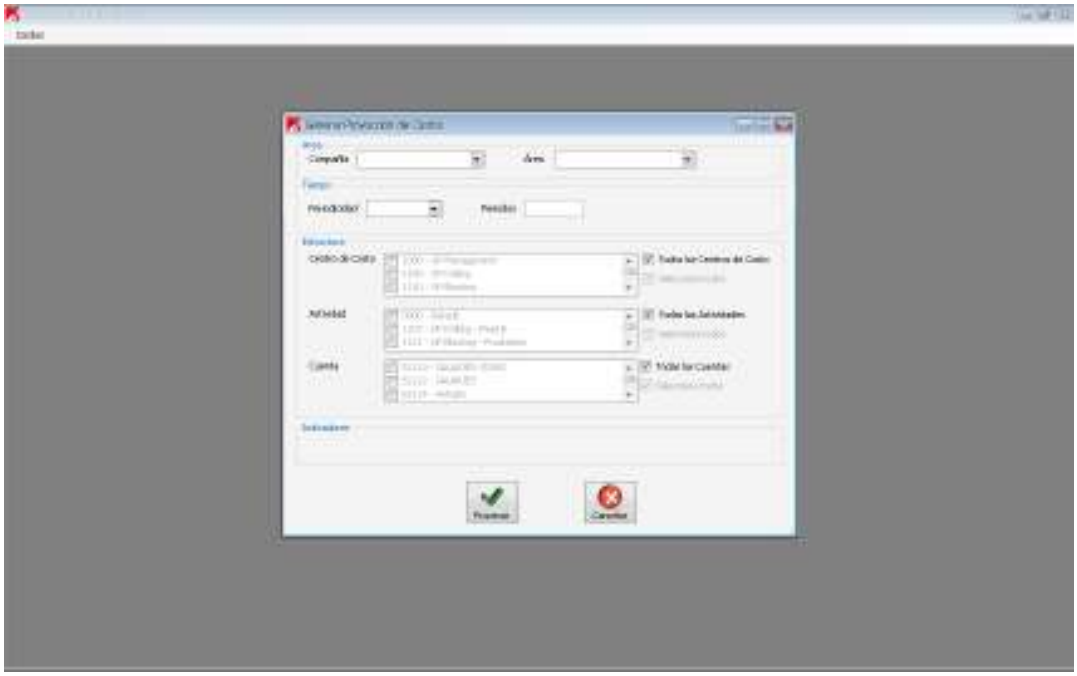


Figura 6.25 Ventana de Proyección de Costos

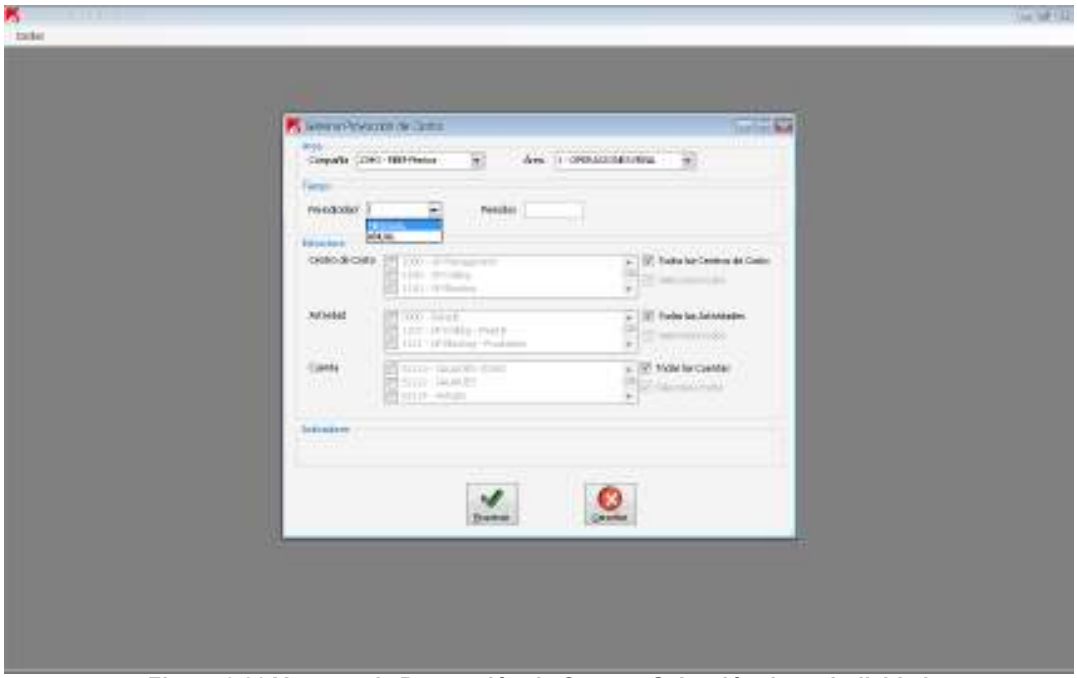


Figura 6.26 Ventana de Proyección de Costos: Selección de periodicidad

Luego de que se ingresan/seleccionan los parámetros, se escoge la opción "Procesar", la cual se va a encargar de procesar el cubo del Datamart de Presupuestos y a su vez actualizar la estructura de minería de datos mediante la dimensión que ésta genera, para luego, a través consultas DMX mostrar la información proyectada de costos. En la Figura 6.27 se muestra un mensaje de confirmación, el cual nos indica

que la información se ha generado correctamente. En caso de no haber información disponible del mismo se informará al usuario.

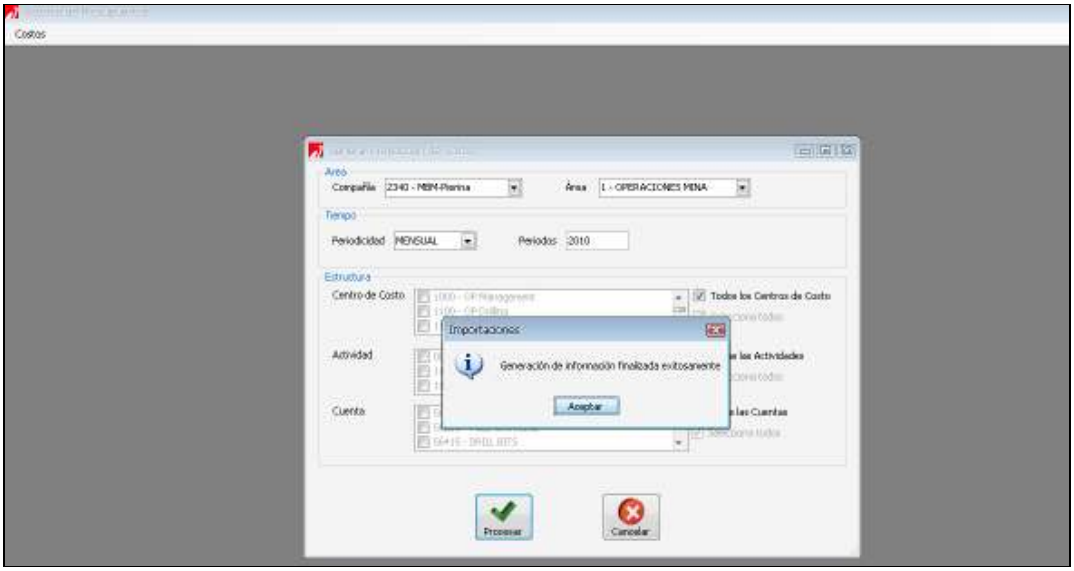


Figura 6.27 Ventana de Proyección de Costos: Confirmación de generación de información

La Figura 6.28 muestra la información proyectada de acuerdo a los criterios seleccionados, la aplicación permite obtener un mayor detalle haciendo búsquedas por ítem, descripción, cuenta, etc.

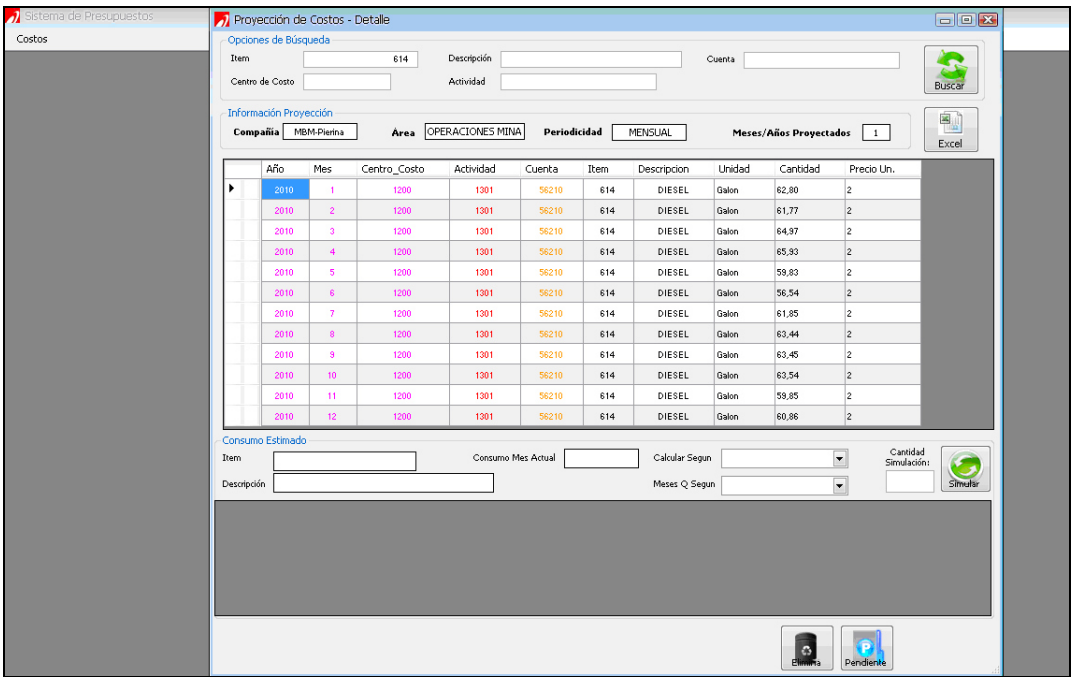
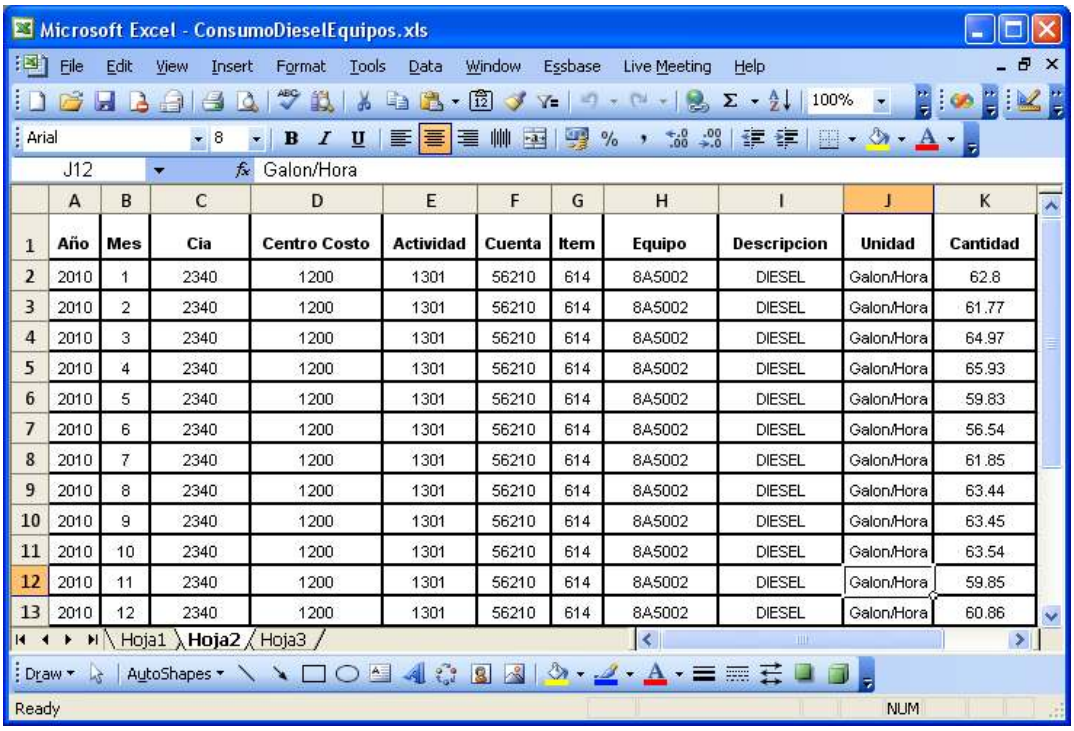


Figura 6.28 Ventana de Proyección de Costos: Visualización de la proyección de costos

La misma información que se muestra en el sistema, puede ser exportada a Excel, tal como lo muestra la Figura 6.29, haciendo click en el botón “Excel” que se muestra en la imagen anterior. Esto es importante ya que los analistas de costos trabajan la mayor parte de sus informes en Excel, y además el ingreso de la información a los sistemas de costos y presupuestación se hacen en base a archivos Excel.



	A	B	C	D	E	F	G	H	I	J	K
1	Año	Mes	Cia	Centro Costo	Actividad	Cuenta	Item	Equipo	Descripcion	Unidad	Cantidad
2	2010	1	2340	1200	1301	56210	614	8A5002	DIESEL	Galon/Hora	62.8
3	2010	2	2340	1200	1301	56210	614	8A5002	DIESEL	Galon/Hora	61.77
4	2010	3	2340	1200	1301	56210	614	8A5002	DIESEL	Galon/Hora	64.97
5	2010	4	2340	1200	1301	56210	614	8A5002	DIESEL	Galon/Hora	65.93
6	2010	5	2340	1200	1301	56210	614	8A5002	DIESEL	Galon/Hora	59.83
7	2010	6	2340	1200	1301	56210	614	8A5002	DIESEL	Galon/Hora	56.54
8	2010	7	2340	1200	1301	56210	614	8A5002	DIESEL	Galon/Hora	61.85
9	2010	8	2340	1200	1301	56210	614	8A5002	DIESEL	Galon/Hora	63.44
10	2010	9	2340	1200	1301	56210	614	8A5002	DIESEL	Galon/Hora	63.45
11	2010	10	2340	1200	1301	56210	614	8A5002	DIESEL	Galon/Hora	63.54
12	2010	11	2340	1200	1301	56210	614	8A5002	DIESEL	Galon/Hora	59.85
13	2010	12	2340	1200	1301	56210	614	8A5002	DIESEL	Galon/Hora	60.86

Figura 6.29 Ventana de Proyección de Costos: Información exportada a Excel para ser utilizada por los analistas de costos

CAPÍTULO 7: Análisis y Simulación de Datos

Luego de presentar el diseño de la solución basado en la técnica de series temporales procederemos a realizar la recolección de datos para de esta manera analizarlos y probar que la propuesta mostrada solucionará el problema en estudio, aumentando la confiabilidad de los índices de los consumibles.

Para la obtención de la muestra se considera el ratio *consumo horario de diesel* en un equipo minero. Este proceso nos ayudará a probar la importancia del aporte a la solución planteada.

7.1 Recolección de Datos

Los datos son recolectados desde los repositorios de datos del sistema implantado, tomando en cuenta sólo el consumo horario de diesel de un equipo minero, para la muestra elegimos al Cargador Frontal Komatsu WA1200 #1, código 8A5002, con periodicidad mensual, para un período de tiempo desde Enero del 2007 hasta Diciembre del 2009.

Por lo tanto, la muestra está formada por 36 meses, se considera como variable continua X al consumo horario de diesel del equipo 8A5002, como se indica en la Tabla 7.1:

Período mensual	X (consumo horario de diesel)
1	62.81771
2	62.23553
3	66.58022
4	61.27045
5	65.77317

6	65.92947
7	63.88638
8	63.81969
9	61.11587
10	57.32908
11	59.71080
12	56.53860
13	57.56056
14	60.59420
15	56.78443
16	54.13467
17	58.09089
18	53.82442
19	63.55317
20	63.55317
21	63.55317
22	63.55317
23	63.55317
24	63.55317
25	59.55189
26	67.86663
27	59.91908
28	62.29160
29	66.03747
30	83.00944
31	106.84350
32	63.83800
33	65.18251
34	62.25712
35	57.74900
36	59.72555
P (Promedio)	63.43297
Períodos con 53 < X <= 56	2
Períodos con 56 < X <= 59	6
Períodos con 59 < X <= 62	10
Períodos con 62 < X <= 65	11
Períodos con 65 < X <= 68	5
Períodos con 68 < X <= 107	2

Tabla 7.1 Tabla resultado del consumo de diesel mensual en el Cargador Frontal WA1200-1

7.2 Simulación de Datos aplicando el Método de Montecarlo

Luego de haber realizado la recolección de datos utilizaremos el algoritmo de Montecarlo para obtener una muestra significativa de lo visto en la sección anterior.

En la Tabla 7.2 convertimos las variables continuas obtenidas en variables discretas, como observamos a continuación:

Ratio mensual de diesel Variable Continua	Ratio mensual de diesel Variable Discreta
54 <= X < 57	54
57 <= X < 60	57
60 <= X < 63	60
63 <= X < 66	63
66 <= X < 69	66
69 <= X < 108	69

Tabla 7.2 Conversión de variables continuas a variables discretas

El método de Montecarlo se basa en la generación de múltiples números aleatorios. El primer paso es determinar la variable aleatoria, en nuestro caso es el ratio de consumo mensual de diesel, con esto obtenemos la frecuencia relativa y la acumulada, tal como se muestra en la Tabla 7.3:

Ratio mensual de Diesel	N° Períodos que tienen ese ratio mensual	Frecuencia Relativa (Total/N° Períodos)	Frecuencia Relativa Acumulada
54	2	0.056	0.056
57	6	0.167	0.222
60	10	0.278	0.500
63	11	0.306	0.806
66	5	0.139	0.944
69	2	0.056	1.000
Total	36	1	

Tabla 7.3 Distribución de Frecuencias obtenidas de la Tabla 7.1

El siguiente paso es generar la tabla de variables aleatorias, calcular el tiempo promedio, varianza, desviación estándar y el error.

Período	Números Aleatorios	X (consumo horario de diesel)	(X – P)	$(X – P)^2$
1	0.966	69	7.68	58.98
2	0.747	63	1.68	2.82
3	0.731	63	1.68	2.82
4	0.461	60	(1.32)	1.74
5	0.395	60	(1.32)	1.74

6	0.228	60	(1.32)	1.74
7	0.743	63	1.68	2.82
8	0.812	66	4.68	21.90
9	0.590	63	1.68	2.82
10	0.680	63	1.68	2.82
11	0.871	66	4.68	21.90
12	0.284	60	(1.32)	1.74
13	0.348	60	(1.32)	1.74
14	0.363	60	(1.32)	1.74
15	0.446	60	(1.32)	1.74
16	0.485	60	(1.32)	1.74
17	0.529	63	1.68	2.82
18	0.377	60	(1.32)	1.74
19	0.118	57	(4.32)	18.66
20	0.735	63	1.68	2.82
21	0.048	54	(7.32)	53.58
22	0.540	63	1.68	2.82
23	0.169	57	(4.32)	18.66
24	0.795	63	1.68	2.82
25	0.646	63	1.68	2.82
26	0.319	60	(1.32)	1.74
27	0.696	63	1.68	2.82
28	0.430	60	(1.32)	1.74
29	0.401	60	(1.32)	1.74
30	0.962	69	7.68	58.98
31	0.534	63	1.68	2.82
32	0.349	60	(1.32)	1.74
33	0.737	63	1.68	2.82
34	0.706	63	1.68	2.82
35	0.034	54	(7.32)	53.58
36	0.186	57	(4.32)	18.66
37	0.251	60	(1.32)	1.74
38	0.543	63	1.68	2.82
39	0.257	60	(1.32)	1.74
40	0.558	63	1.68	2.82
41	0.919	66	4.68	21.90
42	0.166	57	(4.32)	18.66
43	0.702	63	1.68	2.82
44	0.659	63	1.68	2.82
45	0.023	54	(7.32)	53.58
46	0.742	63	1.68	2.82
47	0.833	66	4.68	21.90
48	0.374	60	(1.32)	1.74
49	0.516	63	1.68	2.82
50	0.379	60	(1.32)	1.74
51	0.402	60	(1.32)	1.74
52	0.287	60	(1.32)	1.74
53	0.047	54	(7.32)	53.58
54	0.885	66	4.68	21.90
55	0.768	63	1.68	2.82
56	0.245	60	(1.32)	1.74
57	0.391	60	(1.32)	1.74
58	0.007	54	(7.32)	53.58
59	0.191	57	(4.32)	18.66
60	0.065	57	(4.32)	18.66
61	0.405	60	(1.32)	1.74
62	0.234	60	(1.32)	1.74
63	0.188	57	(4.32)	18.66
64	0.800	63	1.68	2.82
65	0.863	66	4.68	21.90
66	0.404	60	(1.32)	1.74
67	0.026	54	(7.32)	53.58
68	0.820	66	4.68	21.90
69	0.543	63	1.68	2.82
70	0.874	66	4.68	21.90
71	0.895	66	4.68	21.90

72	0.888	66	4.68	21.90
73	0.604	63	1.68	2.82
74	0.402	60	(1.32)	1.74
75	0.287	60	(1.32)	1.74
76	0.595	63	1.68	2.82
77	0.752	63	1.68	2.82
78	0.598	63	1.68	2.82
79	0.489	60	(1.32)	1.74
80	0.481	60	(1.32)	1.74
81	0.163	57	(4.32)	18.66
82	0.590	63	1.68	2.82
83	0.976	69	7.68	58.98
84	0.588	63	1.68	2.82
85	0.685	63	1.68	2.82
86	0.330	60	(1.32)	1.74
87	0.970	69	7.68	58.98
88	0.672	63	1.68	2.82
89	0.550	63	1.68	2.82
90	0.233	60	(1.32)	1.74
91	0.708	63	1.68	2.82
92	0.177	57	(4.32)	18.66
93	0.203	57	(4.32)	18.66
94	0.246	60	(1.32)	1.74
95	0.516	63	1.68	2.82
96	0.811	66	4.68	21.90
97	0.763	63	1.68	2.82
98	0.036	54	(7.32)	53.58
99	0.131	57	(4.32)	18.66
100	0.464	60	(1.32)	1.74
Total		6,132	-2.84E-14	1,211.76
P (Promedio)		61.32		
S^2 (Varianza)		12.12		
S (Desviación Estándar)		3.48		
Error = K*S/Raíz(n)		1.75		

Tabla 7.4 Resultado de la Simulación de Montecarlo

Luego de observar los resultados de 100 iteraciones con el método de Montecarlo, en la Tabla 7.4, observamos que el ratio promedio es de 61.32 galones/hora de diesel, lo cual es menor al ratio promedio de 63.43 galones/hora de diesel mostrado en la Tabla 7.1. Con un error de 1.75 galones/hora de diesel para el equipo 8A5002.

7.3 Simulación de Datos haciendo uso del Sistema de Proyección de Costos

En la Tabla 7.5 podemos observar el resultado de la simulación haciendo uso de la aplicación implementada en este trabajo, el sistema de Proyección de Costos, para el cargador 8A5002 se ha obtenido un ratio de consumo horario de 62.07 gal/hr como promedio en el año 2010, con una desviación estándar de 2.56, la cual es menor a la encontrada usando el método de Motecarlo (3.48).

Año	Mes	Cia	Centro Costo	Actividad	Cuenta	Item	Equipo	Descripcion	Unidad	Cantidad
2010	1	2340	1200	1301	56210	614	8A5002	DIESEL	Galon/Hora	62.80
2010	2	2340	1200	1301	56210	614	8A5002	DIESEL	Galon/Hora	61.77
2010	3	2340	1200	1301	56210	614	8A5002	DIESEL	Galon/Hora	64.97
2010	4	2340	1200	1301	56210	614	8A5002	DIESEL	Galon/Hora	65.93
2010	5	2340	1200	1301	56210	614	8A5002	DIESEL	Galon/Hora	59.83
2010	6	2340	1200	1301	56210	614	8A5002	DIESEL	Galon/Hora	56.54
2010	7	2340	1200	1301	56210	614	8A5002	DIESEL	Galon/Hora	61.85
2010	8	2340	1200	1301	56210	614	8A5002	DIESEL	Galon/Hora	63.44
2010	9	2340	1200	1301	56210	614	8A5002	DIESEL	Galon/Hora	63.45
2010	10	2340	1200	1301	56210	614	8A5002	DIESEL	Galon/Hora	63.54
2010	11	2340	1200	1301	56210	614	8A5002	DIESEL	Galon/Hora	59.85
2010	12	2340	1200	1301	56210	614	8A5002	DIESEL	Galon/Hora	60.86
									Promedio	62.07
									Desviación Estándar	2.56

Tabla 7.5 Simulación del Sistema de Proyección de Costos

CAPÍTULO 8: Conclusiones y trabajos futuros

En los últimos años, ha crecido a velocidades exponenciales nuestra capacidad para almacenar datos, sin embargo, nuestra capacidad para procesarlos y utilizarlos no ha ido a la par; por este motivo, la minería de datos se presenta como una tecnología de apoyo para explorar, analizar, comprender y aplicar el conocimiento obtenido usando gran cantidad de datos. Una de las tareas fundamentales de la minería de datos es descubrir nuevos caminos que nos ayuden en la identificación de interesantes estructuras en los datos. En este trabajo se ha desarrollado un sistema de presupuestación que permite proyectar costos aplicando minería de datos. El presente trabajo demostró la factibilidad de seguir ampliando el campo de aplicación de la minería de datos. En este sentido, se ha logrado obtener las siguientes conclusiones:

- El modelo propuesto del proceso de presupuestación permite lograr una mejora en la obtención de los ratios de consumibles, ya que se ha automatizado el análisis estadístico de datos históricos logrando aumentar la confiabilidad de los índices de consumibles, los cuales se obtendrán en un menor tiempo de procesamiento y se evitarán los errores humanos que actualmente se presentan

- Haciendo uso de la minería de datos se ha podido mejorar el proceso de presupuestación en el área de Operaciones Mina, a través de la reducción del tiempo dedicado por los analistas de costos a presupuestar y a través de la integración de las diversas fuentes de datos en un solo repositorio de donde se obtienen los ratios de los consumibles, los cuales luego de ser explotados con herramientas de minería de datos se logran obtener las proyecciones de costos en reportes especializados.

- El uso de algoritmos de series temporales como técnica de minería de datos nos ha permitido obtener resultados más cercanos a la realidad de la

organización. Tomando en cuenta todas las variables involucradas en el proceso de presupuestación del área Operaciones Mina, concluimos que es la técnica más adecuada para resolver la problemática planteada.

Como podemos observar, este trabajo puede diversificarse y extenderse a las demás áreas operativas de la Mina, que realizan un similar proceso de presupuestación y que requieren de análisis estadístico para el cálculo de indicadores de consumo, por lo que se recomienda, para trabajos futuros una adaptación de la solución dada para que abarque a las áreas de Procesos y Mantenimiento.

CAPÍTULO 9: Referencias Bibliográficas

TESIS:

[Cardona05] Cardona, Carlos M., "Utilidad Práctica derivada de aplicar Minería de Datos en algunas Empresas de Medellín", Trabajo de grado para optar al título de Ingeniero de Sistemas, Universidad EAFIT, Departamento de Informática y Sistemas, 2005, Medellín-Colombia

[Martinez03] Javier Martínez de Pisón, *Optimización mediante técnicas de minería de datos del ciclo de recocido de una línea de galvanizado,* Universidad de la Rioja, Doctorado, 2003, La Rioja-España

[Servente02] Magdalena Servente, *Algoritmos TDIT aplicados a la Minería de Datos Inteligente*, Tesis de grado para optar al Título de Ingeniero Informático, Universidad de Buenos Aires, Facultad de Ingeniería, Laboratorio de Sistemas Inteligentes, 2002, Buenos Aires-Argentina

[Vallejos06] Sofía Vallejos, *Minería de Datos*, Universidad Nacional del Nordeste de Argentina, Pregrado, 2006, Corrientes-Argentina

LIBROS:

[Chapman99] P. Chapman, J. Clinton, T. Khabaza, T. Reinartz and R. Wirthz, *The CRISP-DM process model*, Technical Report, CRISPDM Consortium, 1999.

[Michalski98] Michalski, R.S. Bratko, I. Kubat, *Machine Learning and Data Mining. Methods and Applications*. Wiley & Sons Ltd., 1998, EE.UU

[PMBOK04] Project Management Institute, *Guía de los Fundamentos de la Dirección de Proyectos (Guía del PMBOK®)*. PMI Four Campus Boulevard, 2004, EE.UU

[Vercellis09] Carlo Vercellis, *Data Mining and Optimization for Decision Making*. John Wiley & Sons Ltd, 2009, UK

REVISTAS:

[Aluja01] Aluja, Tomás, *La Minería de Datos, entre la Estadística y la Inteligencia Artificial*. Universitat Politècnica de Catalunya, Revista QÜESTIIÓ, vol. 25, 3, p. 479-498, 2001, España

[Fayyad96] Fayyad U.M., Piatetsky-Shapiro G., Smyth P., *From Data Mining to Knowledge Discovery in Databases.* American Association for Artificial Intelligence, AI Magazine, p.37-53, 1996, USA

[Riquelme06] Jose Riquelme, Roberto Ruiz, Karina Gilbert. *Minería de Datos: Conceptos y Tendencias*. Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial. No.29, p. 11-18, 2006, España

[Siebes00] Siebes A., *Data Mining and Statistics*. Cism Courses and Lectures, Nº 408, International Centre for Mechanical Sciences, CISM, p. 1-38, 2000, Holanda

[Witten00] Ian Witten, Frank Eibe, *Data Mining. Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann Publishers, 2000, USA

PÁGINAS WEB:

[Cont01] Observatorio Iberoamericano de Contabilidad de Gestión, *El Proceso Presupuestario en la Empresa*. http://www.observatorio-iberoamericano.org/paises/spain/04.htm, 16 Noviembre de 2009

[Crisp01] CRISP-DM, Cross Industry Standard Process for Data Mining, http://www.crisp-dm.org/Process/index.htm, 02 Enero de 2010

[Edwards01] JD Edwards EnterpriseOne, Oracle, http://www.oracle.com/lang/es/applications/jdedwards-enterprise-one.html, 26 Diciembre de 2009

[Gartner01] James Richardson, Kurt Schlegel, Rita L. Sallam, Bill Hostmann, http://www.gartner.com/technology/media-products/reprints/cognos/vol6/article2/article2.html; 10 de Diciembre 2009

[Gondar01 José E. Gondar, *Metodologías para la Realización de Proyectos de Data Miningi*, http://www.estadistico.com/arts.html?20040426, 12 Diciembre de 2009

[Mantignon01] Randall Mantignon, *An overview of SAS Enterprise Miner*, http://www.sasenterpriseminer.com/documents/WUSS_Papers.pdf, 02 Enero de 2010

[MEM01] Ministerio de Energía y Minas del Perú. *Estadísticas*. http://www.minem.gob.pe, 01 Febrero de 2010

Microsoft01] Microsoft TechNet, Algoritmos de minería de datos (Analysis Services: Minería de Datos), http://technet.microsoft.com/es-es/library/ms175595.aspx, 30 Diciembre de 2009

[Mol01] Luis Molinero, *Análisis de Series Temporales*. http://www.seh-lelha.org/tseries.htm, 22 Noviembre de 2009

[Nadinic08] Mladen W. Nadinic, Data Mining y Data Warehousing, http://www.scribd.com/doc/19855790/Mineria-de-Datos-y-Data-Warehouse, 19 Octubre 2009

[Oracle01] Oracle, Productos y Servicios, *Oracle and Hyperion*. http://www.oracle.com/hyperion/index.html, 20 Octubre de 2009

[Oracle02] Oracle, Productos y Servicios, *Hyperion Planning-System 9*. http://www.oracle.com/global/es/products/appserver/business-inteligence/hyperion-financial-performance-management/hyperion-planning-system9.html, 20 Octubre de 2009

[Runge01] Runge, *Technology Solutions, Xeras Financial Modelling*. http://www.runge.com/en/technology_solutions/xeras, 22 Octubre de 2009

[Sap01] SAP History, http://www.sap.com/about/company/history/index.epx, 28 Diciembre de 2009

[Stat01] StatSoft, *Elementary Concepts in Statistics.* http://www.statsoft.com/textbook/stathome.html, 30 Noviembre de 2009

ANEXO A

A.1 Caso de Estudio

A continuación se describirá nuestro caso de estudio para poder conocer el ámbito en el cual se manejará la solución.

A.2 Minera Barrick Misquichilca (Perú)

Barrick Gold Corporation es la compañía minera de oro más importante del mundo, actualmente cuenta con 27 minas operativas, y muchos proyectos en desarrollo y en etapa de exploración. Barrick está presente en los cinco continentes y es la compañía aurífera con la mayor cantidad de reservas en la industria.

Inicialmente, en 1987 la Compañía creció mediante adquisiciones en América del Norte, con especial mención en la compra de Goldstrike.

Más tarde, en 1994, con la compra de Lac Minerals Ltd, y Arequipa Resources Ltd. en 1996, la compañía se expandió hacia Sudamérica. La compra de Lac Minerals Ltd. dio a Barrick el control de El Indio en Chile y un interés del 40% en el proyecto Veladero en Argentina. Arequipa Resources aportó propiedades de exploración en Perú, incluyendo Pierina.

En ese entonces, la compañía seguía metódicamente las tres estrategias complementarias que marcan su éxito hasta el día de hoy: inversión permanente en exploración y desarrollo; un enfoque de desarrollo basado en distritos para así optimizar reservas en franjas de oro que parecen ser muy buenos prospectos; y adquisiciones y fusiones disciplinadas.

En 1999, Barrick compró Sutton Resources, cuyas propiedades mineras en Tanzania incluían el depósito Bulyanhulu. Aumentó las reservas de oro de 3,8 a 10 millones de onzas en tan solo 18 meses, y la mina empezó su fase de producción en 2001.

De acuerdo con su estrategia de desarrollo a nivel distrital, en 2000 Barrick compró Pangea Goldields Inc., cuyas propiedades para exploración incluían Tulawaka en Tanzania.

La fusión con la compañía minera Homestake, en 2001, fue un paso importante. Agregó minas en América del Norte y del Sur y –lo que era nuevo para Barrick—en Australia. Además, ayudó a sentar las bases para el cambio organizacional del año 2003, cuando pasó de un modelo centralizado a una plataforma descentralizada, conformada por unidades regionales de negocios.

En paralelo, Barrick estaba desarrollando un nuevo paquete de minas: Tulawaka (Tanzania), Lagunas Norte (Perú), Veladero (Argentina) y Cowal (Australia). Las 3 primeras comenzaron a operar en el año 2005 y la cuarta a principios del 2006.

En enero de 2006, Barrick concluyó un acuerdo amigable con Placer Dome, una adquisición cuyos activos complementarios mejoraron posteriormente la posición de la empresa en América del Norte, Tanzania y Australia; agregaron activos cupríferos de clase mundial en Chile y ampliaron su presencia global a Papúa Nueva Guinea y Sudáfrica. Además, prácticamente duplicaron el tamaño de su equipo a nivel mundial.

Esta transacción es el ejemplo más reciente del enfoque consistente de Barrick, orientado hacia el éxito de sus negocios: adquirir activos de calidad, valorar y formar personas al igual que yacimientos, seguir creciendo y prosperando en un sector demandante.

Minera Barrick Misquichilca es la compañía peruana que maneja los activos de Barrick Gold Corporation desde 1996.

A.3 Visión

Ser la mejor compañía productora de oro del mundo, a través de la exploración, adquisición, desarrollo y producción de reservas de oro de calidad, de manera segura, rentable y socialmente responsable.

A.4 Valores de Oro

Comportarse como Dueños

Aceptamos la responsabilidad de nuestras acciones y de sus resultados. Manejamos los activos de la Compañía como propios. Somos emprendedores y buscamos oportunidades para hacer crecer a nuestra empresa. Actuamos con integridad - operando según la letra y el espíritu de la ley y del Código de Ética de Barrick.

Actuar con un Sentido de Urgencia

Somos decididos, tomamos la iniciativa y tomamos decisiones difíciles cuando son necesarias. Fijamos las prioridades y actuamos según ellas.

Ser un Miembro del Equipo

Trabajamos siguiendo las prácticas de seguridad de la empresa en todo momento. Respetamos a nuestros colegas y a aquellos con los que nos relacionamos fuera de nuestra organización. Escuchamos a otros para entender y pedimos ayuda. Construimos confianza y celebramos nuestros éxitos. Ayudamos a otros para que mejoren su eficiencia. Promovemos la seguridad y la confianza mutua en nuestras capacidades.

Mejorar Continuamente

Siempre estamos comprometidos a mejorar. Construimos en base a buenas ideas, aprendemos de nuestros errores y desafiamos el status quo. Pensamos con amplitud y tenemos un deseo de tener éxito y de agregar valor a nuestro trabajo.

Entregar Resultados

Tenemos una visión clara hacia donde vamos y de como llegar ahí. Enfocamos nuestros recursos para lograr nuestros objetivos. Prestamos mucha atención al detalle y mantenemos nuestros compromisos. Entregamos resultados.

A.5 Estructura Organizacional

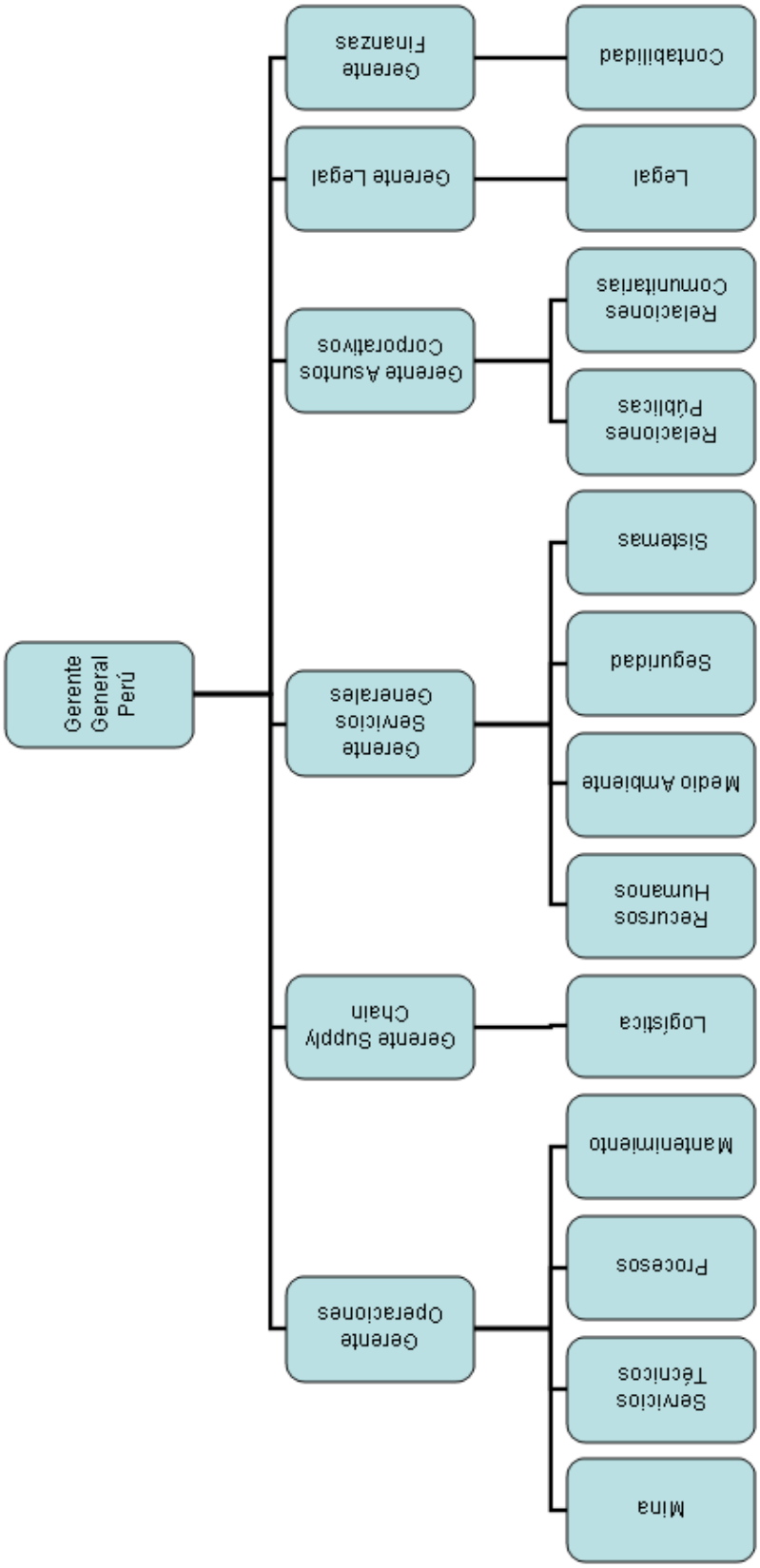


Figura A.1 Organigrama Organizacional de Minera Barrick Misquichilca – Sede Pierina

La Figura A.1 muestra la organización que existe actualmete en Minera Barrick Misquichilca – Sede Pierina, sólo se muestran las áreas principales, ya que cada una se divide en departamentos en el último nivel del árbol organizacional.

Para la implementación de nuestra aplicación nos enfocamos dentro del área de Mina (Operaciones Mina).

A.6 Propuesta de Solución

A.6.1 Situación Actual

En la Figura A.2 se diagrama el actual proceso de presupuestación, donde podemos ubicar como actor principal al analista de costos, ya que esta persona realiza el análisis estadistico de datos que queremos automatizar y optimizar para obtener los indicadores o ratios de consumos de los consumibles principales en el área de Operaciones Mina, tales como: diesel (galones/hora), explosivos (toneladas/taladro), aceros de perforación (metros perforados/hora), principalmente; los cuales se ingresarán en el Sistema Xeras para su procesamiento y costeo, posteriormente se genera un reporte de costos que será ingresado al sistema Hyperion para que esté accesible en línea por los usuarios finales (Gerentes, Superintendentes, Jefes de Area, Analistas de Costos, Contadores, entre otros).

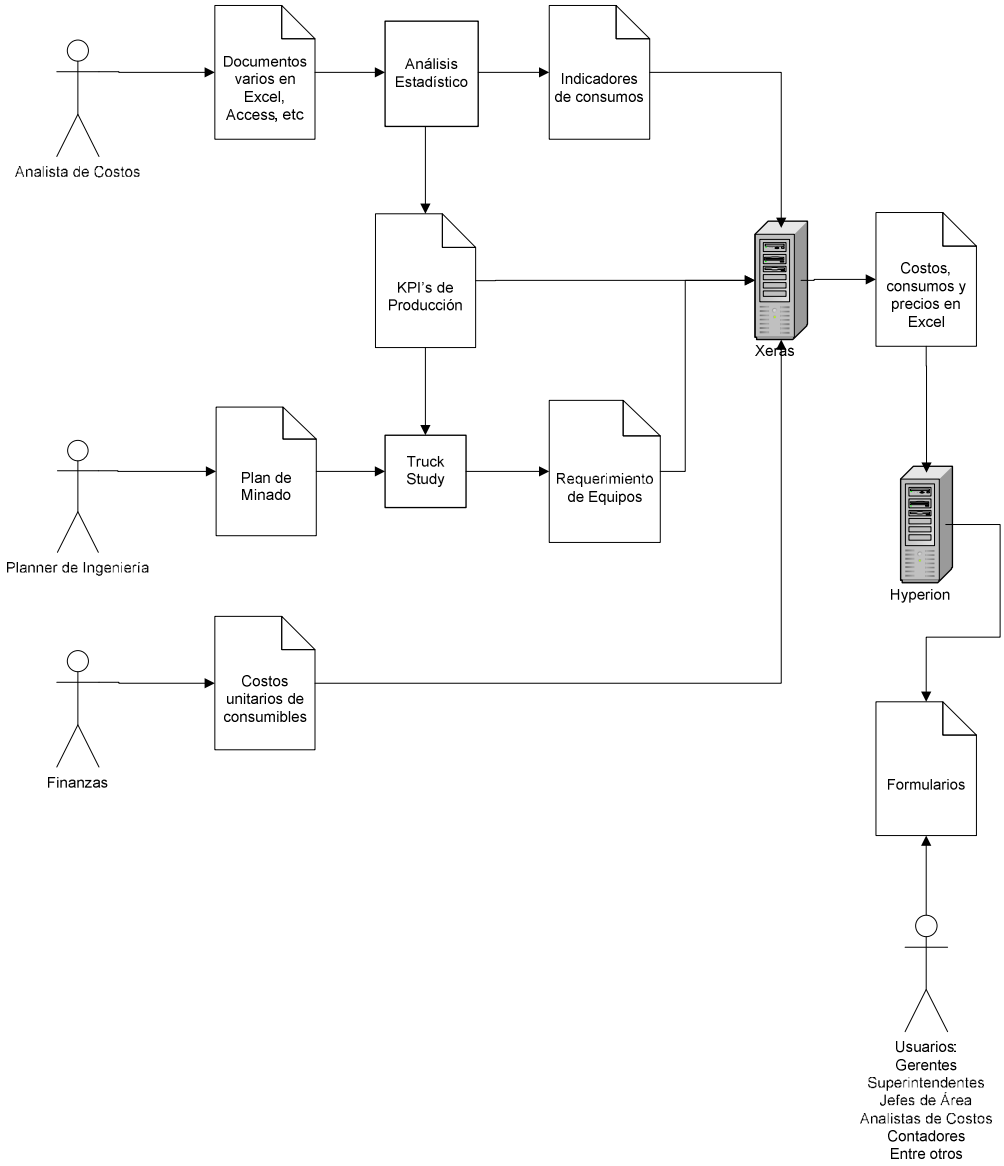


Figura A.2 Proceso Actual de Presupuestación

A.6.2 Situación Propuesta

En la Figura A.3 se diagrama el proceso de presupuestación propuesto, donde podemos observar que el actor principal (analista de costos) accede al Sistema de Proyección de Costos para poder generar ahí los indicadores que servirán de base para el cálculo de los consumibles principales de Mina y continuar con el procedimiento descrito en el punto anterior. El Sistema de Proyección de Costos se conecta al servidor de Base de Datos que es alimentado por proyecciones de costos históricos calculados mediantes series temporales, haciendo uso de datos almacenados en un Datawarehouse provenientes de diversas fuentes.

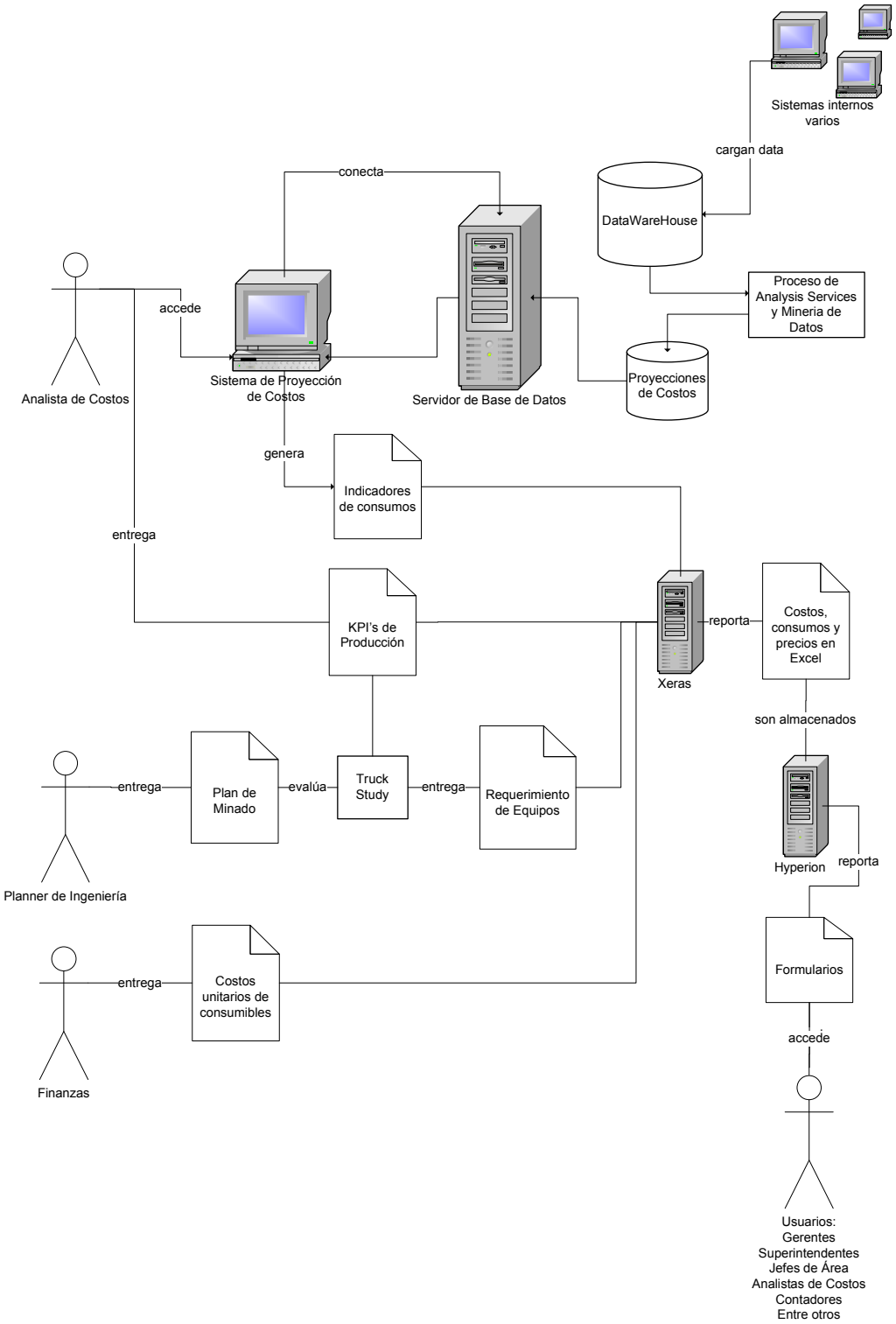


Figura A.3 Proceso Propuesto de Presupuestación